University of Alberta

Rationality, Evolution, and Symbolic Utility

by

William (Guillermo) Innes Barron

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Department of Philosophy

Edmonton, Alberta

Fall 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-68910-7

Canada

University of Alberta

Library Release Form

Name of Author: William (Guillermo) Innes Barron

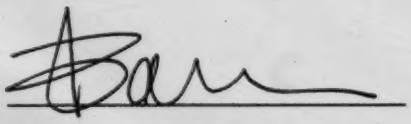
Title of Thesis: Rationality, Evolution, and Symbolic Utility

Degree: Doctor of Philosophy

Year this Degree Granted: 2001

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior permission.



106 - 10656 - 84th Avenue
Edmonton, Alberta,
Canada

14 AUG 2001

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Rationality, Evolution, and Symbolic Utility* submitted by William (Guillermo) Innes Barron in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Wesley Cooper

Dr. Wesley Cooper

Jennifer Welchman

Dr. Jennifer Welchman

Oliver Schulte

Dr. Oliver Schulte

Pamela Willoughby

Dr. Pamela Willoughby

R. P. Morrow

Dr. Ray Morrow

John Baker

Dr. John Baker

Thesis Approval Date: July 21, 2001

Dedication

To my parents, Charles and Sheila Barron, whose love for me I can never adequately return and whose wisdom I have understood far too late in my own life. This work is for them, with much love and affection.

Abstract

This work is a critical analysis and expansion of some of the ideas enunciated by Robert Nozick in his 1993 book The Nature of Rationality. Here Nozick argues that there are strong reasons to expand the classical understanding of rational choice to include three sorts of utility - evidential, causal, and symbolic utility. I offer here an overview of this contention, its implications for rational choice and specific problems within this field, a review of the literature touching on this issue, and then expand Nozick's discussion. I first discuss Dan Sperber's dissent against the widely held view that nonlinguistic symbolic behavior (the object of Nozick's concern) should be understood on the model of language and follow Sperber in arguing for important and fundamental differences between language and other forms of symbolic behavior. I then offer an evolutionary explanation for the widespread existence of symbolic behavior among humans which appeals to its role in resolving recurrent problems of self-announcement and coordination within human societies. I then adapt John Searle's theory of social ontology to explain how symbols attain a social function, and show how that function can be both subjective and objective. I then offer some comments on the role of symbolic actions in intrapersonal and interpersonal (social) processes. By way of an extended appendix, I expand Nozick's employment of evolutionary psychology (EP) to explain features of human rationality and offer an interpretation of EP which is less open to many current objections. I also argue that many accusations of scientific bias which are typically directed at evolutionary psychology are misguided and do not have the epistemic merit which many consider they do.

Acknowledgments

First and foremost, I owe many thanks to my thesis supervisor Wes Cooper for his unflagging support, encouragement, patience, suggestions, and friendship over the many years between this project's inception and its completion. Wes's mentorship has helped me through the Short Runs, the Cold Runs and the Very, Very Long Run.

I also offer many thanks to the members of my committee - Jennifer Welchman (Wes's co-supervisor), Oliver Schulte, Pamela Willoughby, Ray Morrow, and John Baker - for their many helpful suggestions which have in many ways greatly improved this work.

Jean Loates was a constant source of inspiration and encouragement, always reminding me of what was most important in my life, and never allowing me to not believe in myself. I have learned much from her, and without her this work may not have been completed at all.

Mike Pollard and Jon Popowich have been my boon companions for many years. Their help during the course of this project has extended far beyond what I could ever have expected or wished for. They have been as stalwart in seeing me through this scholarly endeavor as they have been on our adventures in the mountains.

Table of Contents

I. Introduction and Overview 1

II. Delving into Rationality 9

Introduction 9

The Idea of Rationality 9

Rationality and Principles 11

The Theory of Rational Choice 16

Game Theory 24

Newcomb's Problem 25

The Prisoner's Dilemma 31

Symbolic Utility 42

What We Still Need to Know About Symbolic Utility 50

III. The Evolution of Symbolic Utility 53

The Nature of Symbolic Action 54

Sperber and the Non-Meaning of Symbols 60

Evolution and the Evocative Power of Symbolic Action 70

The Evolution of Intelligence 75

The Adaptive Value of Symbolic Action 85

A Caveat About Sex Differences 93

IV. The Social Ontology of Symbolic Action 98

Searle's Social Ontology 101

The Function of Function Ascriptions 107

Functions are (Possible) Effects 107

Functions Occur Within Systems 108

Natural and Artifact Function 112
Constitutive and Institutional Rules 126
Status Functions and Language 127
Status Functions: Recursivity, Contestation, and Polemic Force 130

V. Symbolic Action, the Self and the State 138
Principles and Constructing the Self 139
Looking More Closely at Symbolic Action 144
Symbolic Utility and Temporality 147
Symbolic Utility and Politics 148

VI. Evolutionary Psychology 160
Section A: Rationality as Responsiveness to Reasons 160
Reasons and Natural Selection 160

Section B: A Limited Defense of Evolutionary Psychology 167
The Pieces of the Puzzle 175
Some Implications of Evolutionary Psychology 193
Objections to Evolutionary Psychology 195
Some Cases of Special Pleading 195
The Bogeyman of Biological Determinism 202
The Politics of The Politics of Evolutionary Psychology 216
Why Bias Isn't So Bad 225
Sober's Probabilistic Genetic Argument 228
Bias and Probability 237
The Outgroup Bias Effect 240
Works Cited 245

List of Figures

Figure 1: Newcomb's Problem 26

List of Tables

Table 1: The Prisoner's Dilemma 31

Table 2: The Prisoner's Dilemma - Incentive to Cooperate 33

Table 3: The Prisoner's Dilemma - Incentive to Defect 34

Table 4: Independence and Belief Formation 233

Chapter I

Introduction and Overview

We go to gain a little patch of ground
that hath no profit in it but the name.[1]

In the 1981 film Quest for Fire, a small band of hunter-gatherers wander across the Pleistocene landscape desperately seeking any source of fire from which they might rekindle their band's fire. At one point, weary and bedraggled, they pause at the top of a pass to look back at a valley where they have encountered particularly bad luck. One of them leans forward and paws in the ground with his toes, flinging grass and turf behind him towards the valley they have fled, in just the way a dog scratches up grass and dirt to cover its droppings.

No action could more clearly bespeak his disgust and resentment. And yet, in a sense, this is an action that is not meant to communicate anything at all or to anyone in particular. His past tormentors do not see it and his companions, feeling much as he does, learn nothing from seeing him do it. Nor, we may imagine, does he care if anyone sees or understands what he intends. As I hope to demonstrate later, the value of this act lies not in what it communicates, but in what it evokes in the actor's mind. And whatever it evokes in him in some way constitutes a reason for him to perform the act.

Robert Nozick has described an action's property to evoke such feelings as symbolic utility - the value that a symbolic action has independent of whatever

[1] William Shakespeare, Hamlet, (1601) act iv, sc. iv.

its causal, evidential, or linguistic (communicative) value might be.[2] The importance of symbolic actions is indicated by their universality among all human societies and by the sometimes onerous costs people are willing to incur to perform them. We employ symbolic actions to mark almost any act at the center of our personal, social, and religious histories - such rites of passage as birth, the attainment of adulthood, marriage, and death; as well as our pronouncements of friendship, unity, honor, dishonor, investiture, gratitude, war, peace, accomplishment, and so on. And yet as we have noted, what is expressed - what most people call the "meaning" of a symbolic action - may be obvious, uncontroversial, and the action may not even be performed for the purpose of communicating anything in particular to others.

So there are several puzzles here. Why should we care that we perform symbolic actions? Why is symbolic action such a deeply entrenched part of our lives that we barely notice or wonder why we do it? What social forces lead us to maintain a set of symbolic practices? And whence those forces? Given the vast diversity of human social arrangements, what explains the universality of symbolic action? How exactly does a symbolic action fit within a social ontology? And how should we admit symbolic values into our social and political commitments?

Nozick has provided parts of the answers to these questions in The Nature of Rationality (TNOR). My intention in this work is to defend Nozick's project, expand its explanatory scope, and to suggest how it can employed in social and political theory.

---

[2] A symbolic action is simply one which imparts some symbolic utility to its actor, whether or not it has any other utility, communicative or otherwise. Chapter II contains extensive discussion of the concept of symbolic utility.

Chapter II is largely a work of explication in which I lay out Nozick's understanding of rational choice theory and the way in which he embraces symbolic utility within that framework. Two important problems in game theory - Newcomb's Problem and the Prisoners' Dilemma - play a central role in explicating the development of Nozick's conception of decision theory and symbolic utility. I also discuss most of the commentary on TNOR in this chapter.

Chapter III attempts to lay out in greater detail what the implications of human evolution are for evolutionary psychology (EP), and more specifically for the function of symbolic utility within an account of human rationality. Nozick has argued that one role of symbolic action is that it allows an agent to overcome temptation by allowing one undesirable but tempting choice to stand for many others. And Wes Cooper and I have recently argued ("Buridan's Ass", forthcoming) that symbolic utility can usefully solve Buridan's Ass-type problems, in which an agent is thrown into a paralysis of choice because she has no rational reason to prefer one alternative over its twin. But both of these explanations fail to adequately account for three striking aspects of symbolic action: first, symbolic actions are typically not restricted to parametric choice, but more commonly occur within and are made meaningful by a social nexus; second, symbolic action is pervasive to all forms of human life (and neither the temptation-resisting or Buridan's Ass scenarios fully explain why this is is so); and third, that choices fraught with symbolic significance are, more often than not, <u>not</u> choices between eating this piece of cake or not or between apple A or apple B. We do, on occasion, simply <u>assign</u> a symbolic action to some otherwise pedestrian choice that the world presents us but which we might, in other circumstances, have performed in exactly the same way for some quite different reason. But more typically, choosing a symbolic action entails the <u>construction</u> of that choice: it is an option that would not even exist without a desire for symbolic

utility. If one is to marry, one may have to choose between marrying in Calgary or Edmonton, and one may choose to do so for a symbolic reason or not. But the choice to stage an elaborate and ritualized wedding is one that cannot even exist unless we have a desire to express ourselves symbolically.

I therefore contend that the roots of symbolic action must be social ones and I accordingly sketch what I take to be the conditions that shaped human psychology and argue that those conditions support what has been dubbed the Machiavellian hypothesis - the claim that human psychology is largely shaped by the complex and dynamic nature of the social problems that ancestral humans typically encountered in our ancestral environment, and that human psychology has therefore adapted to best serve individual interests in survival and reproduction. This ancestral environment is what is known as our environment of evolutionary adaptation or EEA. For humans, the relevant EEA is the life of hunting and foraging on the African savanna during the Pleistocene that was typical throughout 99% of *Homo*'s history.

I stress that the Machiavellian hypothesis in no way implies that egoism is an essential human trait or that egoism thereby acquires any normative cachet. Symbolic action, I argue, contributes to fitness (roughly, their ability to survive and reproduce)[3] in virtue of its unique potential to allow individuals make sincere self-announcements of commitment to group norms in ways that would have allowed spectators to more reliably and more quickly appraise the reliability of the actor's disposition to cooperate with others. Individuals who engaged in such self-announcing symbolic actions would thereby garner the benefits of the increased likelihood of reciprocal benefits from others and this would increase their fitness.

[3] See Chapter VI, in the section on "Differential Effects on Fitness" for a definition of fitness.

In the second half of chapter III, I argue that one widespread and persuasive model of symbolism, which models symbolism on the structure of language, is mistaken. Here I follow Dan Sperber's account in Rethinking Symbolism with emendation and expansion. As Sperber explains it, there are crucial disanalogies between language and other forms of symbolic behavior that suggest that their sources of motivation cannot arise from similar mental mechanisms. Nor can a linguistic model of symbolism explain the motivation for symbolic action, and this is a central problem in any theory of symbolic action. The human propensity to symbolic action cannot be understood by some one-to-one relationship between signifier and signified, says Sperber, precisely because symbolic action is a response to what is inexplicable and ineffable and that cannot be subsumed into what Sperber calls "encyclopedic knowledge" which can be expressed in spoken language, and it is a mistake to treat symbolic action as such.

Chapter IV attempts to explain how social institutions such as symbolic actions (many of which are, after all, distinctly social in their nature) can arise from the brute facts of physics and biology without recourse to what Daniel Dennett calls "skyhook" explanations (Idea, 74). Here I follow John Searle's account (found in the first chapters of The Construction of Social Reality) of the origins and ontology of social institutions through the ascription of status functions - roughly, those functions which confer causal powers on an entity and which it acquires by a process of collective agreement between members of a social group. However, I argue that Searle's subjectivist account of function ascription here is fundamentally mistaken. The fact that many social objects, in some contexts, acquire their function merely in virtue of some collective act of intentionality is relevant, but simply ascribing a function to something is neither sufficient nor

necessary for that some object have a particular function. Moreover, when observers seek to determine the function of some social entity, they do not assume that an explanation has been achieved simply by finding one person who thinks that F is the function of that entity, and such an account does not explain why claims about the functions of an entity can be either mistaken or contested. Nor will subjective accounts of function always be persuasive or informative in explanations of why an entity exists, why it continues to exist, or why it has the structure it does. I further contend that functional analysis is important in large part because it plays a central role in explanations of this sort. I argue instead for a causal-etiological account of function wherein the analysis of biological, social, and artifact functions is continuous and univocal. However, this approach in no way undermines Searle's treatment of function elsewhere, since intentional properties will frequently (but not always) define social functions and this implication is sufficient to support Searle's account.

Searle further argues that the creation of social institutions via the imposition of status functions is essentially a linguistic matter since there is no way for status functions to be visible to us without some linguistic markers, no matter how crude they are. I concur, but argue that this is only a formal constraint and therefore compatible with Sperber's account of symbolic action that sees the content and evocative power of symbolic action as primarily nonlinguistic attributes.

In chapter V, I discuss how principles, as Nozick construes them, play a role in constructing one's self-conception and the connections between dispositions and principles in human decision making. I then consider how one's allegiance to principles is exemplified in symbolic actions that unite individuals and communities temporally. I then consider some complications that arise for the

arbitration and negotiation of symbolic action within a liberal state.

In Chapter VI, I consider the links between human evolution and human rationality. A good deal of Nozick's explanation of why and how humans are rational relies on his assumption that human rationality is in large part a product of the selective forces at play during our evolution. Nozick defines rationality as a matter of being responsive to reason. Since reasons for and against some belief may be many, conflicting, and multifarious, Nozick suggests this militates strongly for a connectionist model of the mind which weights the values of various reasons. This line of reasoning is pursued to conclude that the domain-specificity of connectionist models provides a powerful motivation to suggest that one useful heuristic with which to investigate why and how minds are organized as they are (and therefore why we have the capacity and desire to perform symbolic actions) is the selective forces that have created the human brain over millions of years. Piecing together an understanding of those forces is an interdisciplinary project involving evolutionary biologists, paleontologists, archaeologists, anthropologists, psychologists, and philosophers, among others.

Evolutionary explanations of human psychology and behavior such as these have been received with little comprehension and less welcome in some quarters, where they are seen as deterministic, reductionist, essentialist, and politically objectionable. Nozick offers little to allay these misunderstands and fears. So in the latter half of Chapter VI, I lay out what I take to be the strongest case for evolutionary psychology (EP) and answer several objections to this research project. Yet even this may not be sufficient: there are numerous methodological assumptions within EP that space simply does not permit me to examine fully, and many individual research programs within EP that I likewise have no space to explain or appraise. And yet the plausibility of EP relies crucially on both a

theoretically sound methodology and on the predictive and explanatory success of particular research programs within its purview. Alas, I can do no more than answer what I take to be the most frequently voiced and most difficult objections to EP and to offer a few illustrative examples of EP's approach, and my chief aim in Chapter VI is to provide some broader theoretical support for the claims in Chapter III. Of course, readers who have no problems with the basic tenets of evolutionary psychology can safely ignore this chapter.

Chapter II

Delving into Rationality

Good reasons must of force, give place to better.[4]

Introduction

The rationale for this chapter is to explicate the major elements of Robert Nozick's account of rationality, decision theory, and symbolic utility that will serve as an entree for the following discussion. Nozick approaches symbolic utility via his conception of rationality and more particularly through his theory of decision value. Although elements of his account of symbolic utility are presaged in his earlier The Examined Life (1990), the bulk of the theoretical support can be found in his 1993 The Nature of Rationality.[5]

The Idea of Rationality

As the title suggests, TNOR is Nozick's attempt to come to grips with some foundational issues in the theory of rationality, but I don't think Nozick intends TNOR as a comprehensive or systematic attempt to explain every aspect of rationality or as an appraisal of all major schools of thought in this complex area. Rather, Nozick seems to want to explore those aspects of rationality that he finds most interesting and to offer provocative suggestions rather than densely argued positions. But, as Christopher Megone has pointed out, while this makes for an

[4] Shakespeare, Julius Caesar (1599), act 4, sc. 3, l. 143.

[5] Unless otherwise indicated, all citations from Nozick's work are from TNOR.

engaging and thought-provoking read, it also detracts from a unified approach to the matter and it is not always easy to see exactly how all the pieces fit together within Nozick's account. Megone argues that the book's chapters (drawn together from three separate lecture series) commence with little explanation as to how they support Nozick's overall conception of rationality, and end with little direction as to how the next chapter will ensue. Moreover, Nozick's own position is not elaborated until halfway through the book (360-1).

I propose in this chapter to explicate Nozick's position on rationality as clearly as I can, to answer some objections against his theory, and to thereby set the stage for a more detailed examination of symbolic utility in the chapters to follow.[6]

Rationality, we might think, is a faculty of the human mind, but note that we also describe beliefs, decisions, desires, goals, actions, arguments, institutions and practices (like science), and even attitudes as "rational." So our account will have to determine whether we are using the term univocally or meaningfully across all these cases and whether one species of usages (that is, the use of "rational" as it applies to one set of cases) is the paradigm by which we can can judge the other usages. Moreover, we employ the concept of rationality as a normative test across a wide spectrum of human endeavors, including economics, logic, moral reasoning, game theory, and political choice to name but a few. These considerations by themselves should warn us that rationality cannot be easily defined or characterized.

Since Immanuel Kant, theorists have traditionally recognized that rationality comprised three distinct elements: cognitive or theoretical rationality tells us

[6] Unfortunately, TNOR has not attracted the critical attention that I think it deserves. In addition to Wes Cooper's "Parfit, My Heroic Death, and Symbolic Utility", the only published commentaries on TNOR are Moser, Christensen, Mellema, and Megone.

what to believe, practical or instrumental rationality tells us what to do, and evaluative rationality tells us what goals we ought to pursue (Rescher Rationality, 3). I consider here only the instrumental aspects of rationality.

Rationality and Principles

Rationality, says Robert Nozick, is a matter of being responsive to reasons, because, as he claims, being responsive to reasons makes it more likely that our beliefs will be true and our actions will satisfy our desires. Reasons themselves are factually connected with hypotheses (about candidate beliefs or actions, say) in such a way that reasons support those hypotheses by increasing their (subjective or objective) probability and humans can also recognize such connections.[7] Moreover, it is no accident that the factual connections between reasons and hypotheses are both evidential and detectable. If the factual and structural connections between particular reasons and hypotheses were stable over long periods of time in the past, and if the capacity to detect those connections contributed to an organism's fitness and was heritable, then natural selection would have favored organisms who could see such connections as "self-evident" (107-8). These reasons will typically be expressed (explicitly or implicitly) as principles.

Roughly, I understand a principle as a generalization which scopes over some set of entities such that it describes or prescribes some feature(s) of them. A principle is nomological if it describes some feature common to all such entities, and normative if it prescribes a certain feature. Nozick's treatment is a bit different.

[7] Nozick doesn't indicate here whether he is referring to subjective or objective probabilities. But discussion in Philosophical Explanations, to which he refers, suggests he may be thinking of subjective probabilities (251ff).

"Principles," Nozick says, "are transmission devices for probability or support that flow from data to cases, via the principle, to judgments and predictions about new observations or cases whose status otherwise is unknown or less certain (5)." Principles do this by grouping various entities under a common rubric and this allows them to perform several important functions (1).

First, principles (such as those found in science, law, and ethics) function intellectually as generalizations about large numbers of cases that allow us to make inferences about new cases, or to understand the the underlying commonality that unites them. In law, for example, the motivation to subsume individual cases under a principle derives its motivation from the sense that like cases should treated alike. Since cases can be alike in many ways, principles direct jurists to the relevant similarities. As in science, legal principles gain support when they cover numerous individual cases, and our decisions in those cases are rendered more acceptable when we can find some principle under which to subsume them. These principles can then be used to make predictions and to inform or persuade others. Principles do this where they possess properties (simplicity, universalizability, predictive power, etc.) that are counted as virtues within their given domain.

Further, an agent who follows a set of principles will benefit those who interact with her because they can reliably predict her behavior and will thus be able to trust her, even in the face of considerable temptation. The agent thus acquires a reputation effect that allows her also to benefit from the increased cooperation of others.

Perhaps, however, Nozick overstates the role of principles here. On the one hand, Nozick's argument presupposes that arguing for the virtues of one

principle over another requires that the principles be explicit if they are even to be understood as principles. And while it is clearly true that in many cases (legal and business dealings) explicit and mutually understood principles may be the only way to regulate interpersonal expectations, it is not clear that this is always the case. Much human interaction (like the use of language, for example) relies on individuals following "rules" and being able to identify violations of those rules, even though they may not be able to make those rules explicit. John Searle argues that the "background" of human thought allows us to create sets of intentionality that are "casually sensitive" to specific rules without internalizing the specific intentionality that expresses those rules (Construction 141-2).

And even members of non-human species can predict and rely on the behavior of conspecifics in complicated ways, even though they are incapable of articulating the principles governing their own or the other's actions. The suggestion here is that much of the interpersonal reliability that Nozick attributes to publicly understood principles could in fact be accomplished by less intellectually demanding measures, and this suggests that there may be a deeper reason why we use principles (9-12). That is, if principles are not needed to interpret others, their appeal may lie elsewhere.

Even though it may be difficult to determine just what principle(s) another is following (because the person acts inconsistently or is deceitful about her principles), knowing those principles nonetheless allows us to predict and interpret the other other's actions because principles also fulfill the personal function of shaping (at least in part) an individual's own behavior.[8] But why

[8] Nozick's initial description of principles embraces both normative (legal, moral, and prudential) and descriptive principles. But in explicating their inter- and intrapersonal roles, he seems only to be concerned with normative principles.

should principles direct our actions? One suggestion might be that to do so would make us consistent, but Nozick, who interprets "consistency" quite narrowly as logical consistency, rejects this notion on the grounds that it is needless to avoid logical inconsistency in our actions since it is impossible to be act in a way that is logically inconsistent (13). Rather, Nozick thinks we follow principles because they increase internal coherence and integrity of our chosen selves - its "organic unity" in Nozick's favored phraseology. But this doesn't seem to me to be clearly right if we interpret consistency a bit more liberally. That is, people may align their actions with their chosen principles because to do otherwise renders their actions inconsistent (in the sense of self-frustrating) with each other and with their goals. So, for example, it would be inconsistent, in this broader sense, to curry favor with one's employer one day and insult her the next. There is no need - yet - to invoke any quasi-mystical sense of organic unity to explain why one might follow principles.

Less controversially, Nozick also notes that following principles can reduce our calculation time in decision making and help us overcome temptation. That is, adherence to a principle can help an agent avoid actions which, at some time T1, she wishes to avoid doing at some later time T2 (when she will be tempted to perform the action), and which she will subsequently regret at T3. In part, this is because the agent may believe she will be more likely to perform such undesirable acts in the future, and in at least some cases, the fact that an agent can avoid the action at T2 is because she believes that the action in some way symbolizes future actions of a similar sort, and this somehow "forges" a connection between this action and future similar actions (26). How symbolic work of this sort is possible, and the intangibles involved, is of course a large part of what this work is about. In chapter III, I raise doubts about this interpretation, and offer an alternative account of the central purpose of symbolic

action.

But what then is the real function of principles within human cognition? This work is not primarily about principles, and I therefore am reluctant to offer any definitive answer. One possibility may lie in the realization that principles derive most of their force from their status as linguistic entities. This claim does not imply that principles exist only as linguistic entities, that is, as part of a shared language that in some way shapes the content and structure of our thoughts: perhaps what Stephen Pinker dubs "mentalese" (Mind, 69-70, 86-90), the language of thinking, is naturally attuned in some way to subsuming knowledge under principles. But such subsumption, even if it is conscious, need not be explicit in that we do not need to clearly formulate to ourselves, in many cases, the exact principle under which we are acting or deciding. To the extent that principles are linguistic entities, they derive their value from their communicative , normative, and polemical effects on others. If Nozick is right in claiming that principles serve to transfer evidential support from one class of cases to a larger class, and if principles show themselves most clearly as linguistic entities - and since language is paradigmatically a social institution - then the central role and purpose of principles may be interpersonal, rather than intrapersonal.

The Theory of Rational Choice

The theory of rational choice plays three interrelated roles within the explanatory framework of TNOR. First, rational choice theory can be used to predict which of a set of strategies an organism should employ to maximize fitness. Since this is so, organisms who followed those strategies (even if they did so without any explicit awareness or intention) would have increased their numbers relative to the increase in numbers of conspecifics who did not follow such strategies. In this way, dispositions for any behavior favored by rational choice theory would have become species typical, and many forms of animal behavior can be interpreted in this way. Thus a large part of the explanation for the existence of human rationality lies in recognizing how it has contributed to solving specific problems that recurred during our distant past, rather than as a domain-general capacity for abstract thought.

(As an aside, W. S. Cooper argues that the theory of rational choice, as it has been classically understood, suffers a severe flaw, and that we can better understand rational choice as a subset of adaptive evolutionary theory. Cooper poses the following two problems which he says demonstrate these two facts. I have omitted the calculations to save space.

(1) Suppose you must bet on the outcome of a perfectly random draw from an urn containing 60 white balls and 40 black balls. You will double your money if you correctly predict the draw, and lose three-quarters of your bet if you bet incorrectly. Moreover, you can make more than one bet, so it is possible to hedge your bet by betting on white and black. Imagine next that you must pick between two betting strategies: (A) bet everything on white for an expected return of about 1.30 times the original stake, or (B) bet 5/8 on white and 3/8 on black for

an expected return of about 1.17. Clearly, on average, one will do better to employ strategy (A) (461).

(2) The situation is exactly as those described in (1) above, but the betting cycle is repeated with the winning from each round serving as the bet on the next round. And this continues for one thousand bets. And, as in (1) you can choose between strategies Since, as classical decision theory demands, a rational agent will always make the same choice in identical situations (excepts in some cases of interactive choice, where randomization is necessary to prevent an opponent from benefiting by being able to predict one's behavior), it seems that here too you should choose strategy A since it maximizes your expected payoff on each round.

But this, surprisingly enough, turns out not to be so. In fact, if you play A for 1000 successive bets, your original stake will be whittled away to an infinitesimal amount. On the other hand, if you play B, you will become fabulously wealthy - on the order of 10^{60} times your original bet. And there is only a miniscule (2×10^{-28}) chance that strategy A will ever pay off better than B (462).

According to Cooper, it is not difficult to find at least theoretic examples of this surprising fact in evolutionary strategies. Imagine that a given species will double its population if its coloration matches ground color (white if there's snow, black if there's no snow), but will lose three-quarters of its population to predation if its coloration is inappropriate relative to the snow cover or lack thereof. And suppose further that there is a 0.6 chance each winter that there is snow. Since this example is an exact analogy of the ball drawing example in (2), we have here a case in which there is an adaptive strategy which is wildly

successful in certain conditions - but which is proscribed by classical decision theory (469).

After replying to obvious objections, Cooper says there are four possible resolutions to this apparent contradiction (i) this oddity is subsumable under classical decision theory, (ii) biological explanations are descriptive, but the theory of rationality is normative and hence no revision is necessary, (iii) the effects are too weak to be significant in the real world, and hence irrelevant, and (iv) "the traditional theory of rationality is invalid as it stands and is in need of biological repair (479)." Cooper opts for position (iv).

But I do not see that Cooper has given any reason to accept this conclusion. In the first place, notice that Cooper has not offered a counterinstance to rational choice theory which is drawn from real world evolutionary biology. Rather, he has adapted a paradoxical case which arises in probability theory to an imaginary (but not improbable) situation in evolutionary adaptation. And, along the way, Cooper has to make a large number of caveats to even make the example biologically acceptable (467-72) and to define the problem in such a way that fitness maximization is in fact equivalent to being rational. If anything, Cooper's point is made much more effectively , and without the necessity of much biological nuance, in the ball-choosing example.

This suggests that there is in fact nothing whatsoever in Cooper's biological example to support the theoretic, heuristic, or methodological supremacy of evolutionary biology over the the theory of rational choice. The very fact that Cooper initially lays out the problem (in a much clearer fashion and with fewer theoretical complications) in a non-biological example proves this is so. The paradox is dissolved once we notice that the cases described by (1) and (2) are

not identical and this means that an agent need not employ the same strategy in both cases. Although strategy A does maximize average returns on a single bet, it does not minimize losses. In the case a black ball is drawn, your bet is reduced to 0.25 its original value. This need not deter you (even if you are somewhat risk adverse) since you can be confident that, over the long run, employing this strategy on many independent bets pays off better than B.

But this is only the case if your bet on some given round does not depend on the winnings on the last round, as (2) specifies. Here the weaknesses of A are apparent. If you lose on any given round, the three-quarter loss will require two successive wins to recoup. On the other hand, if you employ strategy B, you will only lose 3/32 of your bet any time that black comes up (Cooper, 461) and this smaller loss will be easier to win back. So, in cases of bet-multiplying, probability-weighted arithmetic means (as to used to defend strategy A in the single draw) do not accurately reflect the iterated costs of large losses, while a geometric mean does. And all this is noted by Cooper, who argues that this phenomenon is well-known in investment portfolio theory (462-3). Given these considerations, it is difficult indeed to see why Cooper or anyone else should see this paradox as a peculiarly biological one, or why it is a challenge to standard decision theory at all.)

Second, rational choice theory offers insights into the actions of individual humans. This is not to say that humans are always and everywhere rational, but that at least some of their actions can be understood as following the dictates of rational choice theory. Third, much collective social action can only be understood as the aggregative consequence of individual actions. Many social explanations are causal explanations, but the fact that human beings are intentional means that social causal explanations differ importantly from other

causal explanations. The theory of rational choice explains the links between aspects of human intentionality (such as purposes, beliefs, desires, etc.) and the actions that they generate (Little, 39).

It may then be helpful to offer a brief overview of the central tenets of the theory of rational choice, and I have here followed Daniel Little's very accessible exposition in Varieties of Social Explanation. On Little's account the theory of rational action makes two assumptions:

1. Humans have goals that they wish to achieve, and which theorists typically express as their preferences for certain outcomes. Rational choice theory is "thin" in that it says little about the content of these goals. The economic theory of rationality is largely an instrumental one. Economists and many political philosophers often further assume that agents are egoists but this assumption is in no way essential to the theory.

2. Human beings decide what strategy to pursue to achieve their goals based on their beliefs about:

    a. the options available to them and
    b. the probable consequences (costs and benefits) of each choice.

To be rational, an agent need not possess perfect knowledge about her options and their consequences, but she must (at least on some accounts) obtain her beliefs in some rational way (41). An agent, of course, does not typically have just one goal. Typically she will have many goals, which will be of differing value to

her and often heterogeneous and competing. An agent's preference ordering ranks the differing values of goods ordinally, but does not specify the magnitude of difference between these values. We can refine this by assigning a utility function to various outcomes that represent cardinally the relative values of those outcomes. In Little's words, such a theory requires

> a. that utility is a function that takes outcomes as a variable and specifies the value of the good to the agent as a result,
> b. that a rational agent always prefers outcomes with greater utilities, and
> c. that the utility scale is continuous, i.e., it is possible to add utilities (45).[9]

Moreover, a rational agent will, in principle, be able to assign a probability function to each outcome. Probabilities express risk (the relative frequency of desirable outcomes compared to all outcomes for a given action) and uncertainty (our degree of warrant in asserting what those relative frequencies are).[10] Probabilities can be objective or subjective.

Once an agent has assigned utility (U) and probability (P) functions to all outcomes (O) for all choices (C), she can then calculate the utility value for all choices (C_1 ... C_N) as the sum of the products of each outcome and its probability, thus

[9] This is to avoid the problem of lexicographical ordering of entities under which entities are rated first by their relative standings on one criterion, and then by others. So, for example, a nation that wins two gold Olympic medals and no silver medals ranks lexicographically above another nation that wins only one gold medal but fifty silver medals. See Heap et al Theory of Choice, 330-2, for discussion.

[10] The distinction, dating to J. M. Keynes, is a useful one. If, for example, I know that 1% of all automobile tires are likely to fail, I know the risk of buying a tire. But if I do not know what frequency of tires will fail, then I have to make my decision to buy a tire or not under uncertainty. See Heaps et al Theory of Choice, 349-50, for discussion.

1. $UC_1 = P(O_{1,1}) \text{ X } U(O_{1,1}) + POW \text{ X } U(O_{1,2})...$

2. $UC_2 = P(O_{2,1}) \text{ X } (UO_{2,1}) + P(O_{2,2}) \text{ X } U(O_{2,2}) ...$

...

3. $UC_N = P(O_{N,1}) \text{ X } U(O_{N,1}) + P(O_{N,2}) \text{ X } U(O_{N,2}) ...$

Such an approach expresses utilities but does not directly tell the agent how to choose between choices, To do this, the agent must employ a decision rule that directs her how to interpret utility calculations. Three rules commend themselves here. The expected utility rule suggests that the agent ought to pick the choice with the highest probability-weighted utility, on the assumption that to do so will result in the largest expected utility when applied over a large set of cases. The maximin rule, by contrast, notes that some choices with a higher expected utility may nonetheless have disastrous outcomes, and therefore recommends the choice that has the best worst outcome. To see how this is so, suppose that some choice C1 has a higher expected utility than its alternative C2, but carries with it also the chance of a very undesirable outcome. C2, on the other hand, has a lower overall expected utility, but none of its possible outcomes are especially unfortunate. The utility maximizer would therefore choose C1, while the more risk-averse[11] agent would choose C2.

But an agent may reject both these approaches on the grounds that the costs of gathering the requisite information to make either utility-maximizing or risk-avoiding decisions are too high. She may then opt for a satisficing decision rule that simply demands that she pick the first available choice that fulfills some minimal level of utility (Little, 49-51). But this appears paradoxical: given that

[11] That is, an agent who will not accept an actuarially fair gamble. For example, she would not spen $10 to buy a 1 in 1000 chance to win $10000 (Heap et al Theory of Choice, 350-1).

one pursues utility (and that utility is defined by the fact that it is what we pursue), why would we count it as rationally defensible to settle for less? But the objector can argue that the paradox lies instead with utility maximizers, and that it is in fact more rational to behave irrationally (Gauthier Morals, 184). That is, human beings may be natural satisficers simply because it is too costly to calculate which decision will maximize utility. So it may be more cost-effective to not calculate, and to settle for a non-maximizing option. But, David Gauthier argues, although it may be rational to satisfice under conditions of imperfect knowledge where calculation costs are high, it would not be rational to eschew the maximizing action if someone presented it to us free of costs (Gauthier, 186). David Krepps notes further that if one calculates that the costs of calculating the utilities of various options is greater than the expected gain from optimizing then one has an optimizing reason to satisfice (Game Theory, 180). So an apparent case of satisficing may only be a local decision made within a global optimization (or maximization) strategy. In any event, since in many problems of interest, people do wish to maximize utility, decision theorists frequently do assume that utility maximization is the rational agent's concern. But decision theorists need not be wedded to this assumption as a descriptive theory of human psychology across the broad range of human behavior. And since, so far as I can see, my intentions do not rely heavily on assuming humans typically maximize or satisfice in any given situation, I can safely avoid the question. I shall, however, return to the question in chapter VI, where the question of maximization vs. satisficing is relevant to the question of fitness.

Game Theory

A good deal more can be said to further formalize this account, but the preceding will suffice to explicate the basic elements of parametric rational choice for a single agent where outcomes are not affected by the choices of other rational decision makers. In cases of strategic rationality, the outcomes of choices are affected by the decisions of other agents and the situation accordingly becomes much more complex. Decision theory treats such cases as games that are governed by rules and that have two or more players. Players then may or may not have complete information about the rules of the game, the payoffs for various moves, and the strategies available to the other player(s).

Although game theory is in many ways almost impossibly demanding in its requirements for information and rationality, and fails to model the complexities of real world social interaction, it is nonetheless possible to use it to develop aggregative explanations of social behavior. That is, since large scale social behavior is the aggregate of the actions of numerous rational agents, social decisions can be seen as the product of numerous individual rational choices. Of course, it in no way follows that social features exist because individuals actively choose them (Little Theory, 42). Many social outcomes may not be intended by their participants, who may be pursuing very different goals. The free market, for example, frequently succeeds in bringing supply and demand into equilibrium, even though buyers and sellers are only pursuing their own interests.

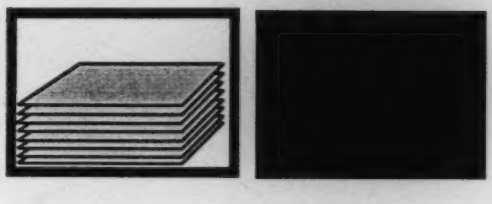
There are two classic problems in game theory that will serve to demonstrate how game theory works and that Nozick uses to explicate and defend his

proposed revisions to rational choice.

Newcomb's Problem

The first is Newcomb's Problem (NP), which Nozick describes thus:

> A being in whose power to predict your choices correctly you have great confidence is going to predict your choice in the following situation. There are two boxes, B1 and B2. Box B1 [which is transparent] contains $1,000; box B2 [which is opaque] contains either $1,000,000 ($M) or nothing. You have a choice between two actions: (1) taking what is in the both boxes; (2) taking only what is in the second box. Furthermore, you know and the being knows you know, and so on, that if the being predicts you will take what is in both boxes, he does not put the $M in the second box; if the being predicts you will take only what is in the second box he does put the $M in the second box. First the being makes his prediction; then he puts the $M in the second box or not, according to his prediction; then you make your choice (41).



B1 B2

$K $M or $0

Figure 1

There are two powerful arguments in favor of each choice, and each relies on a different conception of what counts as a rational choice:

1. The Two-Box Choice: This view notes that either the Being has placed the $M in B2 or he hasn't. And this matter is fixed, it simply can't change. If the Being has placed the money in B2, your best choice is to take both boxes, since $M + $K is more than $M. On the other hand, if the Being has not placed the money in B2, then you should still take both boxes, thereby pocketing $K rather than nothing. Hence the decision to take both boxes dominates the one-box choice.

2. The One-Box Choice: On the other hand, if you are "greedy" enough to choose both boxes, the Being would almost certainly have predicted this beforehand, and would therefore not have put the $M in B2. But if you choose B2 only, the Being would have foreseen this too, and would have placed the money there. So settling for one box is the best way to ensure that the $M will be inside when you open it.

Still, the two-box faction insists, it is a canon of rational choice that it is always

rational to prefer a dominant choice to any other.[12] And choosing both boxes is clearly the dominant choice: whether or not the Being has placed the money in B2 , choosing B2 and B1 is always better than choosing B1. So, as a rational chooser, you apparently can do no other. Therefore, insofar as you are a rational agent, the being must recognize you as such, will predict that you will choose both boxes, and will leave B2 empty. Accordingly, if (counterfactually and *per impossibile*) you were to choose one box (B2), you would be almost certain to find it empty. Hence, the one-box argument is invalid. To this, the one-boxer replies that an agent who is "irrational" enough to reject the dominant choice and choose one box only (which choice the Being would also have foreseen) will almost certainly find $M inside. So, it seems, if rationality is supposed to be primarily instrumental in value, it fails here, since the "irrational" person will fare better than the "rational" one: "If you're so smart, how come you ain't rich?" (Heap et al, 342)

A third approach, suggested to me by a former student, is to take both boxes, on the argument that if one is truly concerned to improve one's financial position, no matter how little, one shouldn't and needn't bother with deep levels of metaphysical speculation about the likely actions of an almost incomprehensibly prescient being. Instead one should just settle for the certainty of receiving the much lesser sum of $1000. This answer, at least, has the advantage that it is completely convincing for those who are content to satisfice their monetary desires at a certain level. Of course, dropping the assumption that one wants to maximize utility (and is willing to risk losing $1000 to do so) drains Newcomb's problem of much of its theoretical significance (and this, after all, is a problem designed to winkle out our intuitions about causal and evidential utility, not an

[12] In decision theory, a dominant choice is one which does as well as, and at least in one outcome, better, than any other choice available to the agent (Dixit and Skeath, 83-4).

exercise to help us solve our financial problems in real life.)

Since Nozick first described this problem in 1969, it has attracted considerable attention (see for example: Levi, Mackie, Talbott, Sobel, and Hurley), but Nozick did not return to the problem until TNOR. Here Nozick notes that outcomes are frequently conditionally dependent on an agent's actions. That is, knowing that an agent has performed (or will perform) some action may be reason to revise one's estimate of the probability that some outcome will obtain. There are two ways in which this can happen. If an agent's actions cause the outcome to happen then the action has causally expected utility (CEU). We may count this as the orthodox view, and the one that supports dominance. On the other hand, an agent's actions may not necessarily cause the event to happen but are nonetheless evidence that it will occur, in which case they possess evidentially expected utility (EEU). So, for example, certain religionists who believe in the doctrine of election (let's call them Electionists) consider that a believer's material well-being on earth is evidence of her future rewards in heaven. Or suppose again that an individual has a gene that predisposes her to a certain profession and to a certain fatal disease. In each case, one's actions (acquiring wealth or entering the given profession) are evidence of one's fate, but in neither case is it rational to act (i.e., to acquire wealth or to avoid the profession) as a means to effect the desired outcome of salvation or avoiding a fatal disease, because the likelihood of those outcomes is not affected by one's actions. So, it would appear, we should not ordinarily act on the basis of EEU considerations (42, 46).

But Nozick purports to show that those who are immune to the blandishments of evidential utility in Electionist and "career gene" examples may be otherwise inclined when confronted with a Newcomb's problem. Nozick offers three

reasons why this should be so. In the first place, despite considerable commentary, there is no clear consensus on which choice is preferable or why. Second, he contends that most two-boxers, no matter how loyal they may be to causal utility, will be very uncomfortable with their choice should the amount in B1 decrease to some negligible sum (say, one cent). In such a case it may be wiser to choose one box, since in doing so, one stands at least some chance of getting $M, and loses only the almost worthless chance of acquiring a single cent. On the other hand, one-boxers will be similarly discomfited should the amount in B1 rise to just under $M. Here they may justifiably suspect that the certainty of receiving a little under $M (by picking two boxes) is preferable to the near-certainty of getting exactly $M.[13]

However, David Christensen thinks that committed two-boxers (himself included) will not convert to the evidentialist position even when B1 is empty. As Christensen puts it, "taking the second box would cost me no money, and I would gain the satisfaction of not bowing to irrational impulse! (261)" But surely Christensen is begging the question on this point. If indeed the Being knows that Christensen is firmly committed to choosing both boxes, then he has not placed the $M in B2, and Christensen's disposition to choose two boxes is exactly what has cost him the $M. And his satisfaction will therefore be both illusory and short-lived. So the mere fact of Christensen's refusal to be swayed as Nozick predicts one-boxers will be does not count as a strong objection to Nozick since Christensen's reasons for doing so don't appear particularly compelling.

All this, argues Nozick, shows that no-one is completely secure in her adherence to either EEU or CEU. And finally, whether or not we should depend primarily

[13] As Nozick points out, one's decision here should also be influenced here by the degree of faith one has in the infallibility of the being's predictions (44).

on EEU or CEU depends on how many Newcomb-type problems there happen to be in the world. If there are many, EEU will be very important, and if otherwise, less. Accordingly, Nozick proposes that rational choice theory should reflect this nuance via the device of decision value (DV):

$$4.\ DV(A) = Wc \times CEU(A) + We \times EEU(A).$$

where Wc and We are (respectively) the weights attached to the causal and evidential utility of some act A. Importantly, these weights reflect the "legitimate force" of each sort of utility in a given decision and allow an agent to shift her preference in Newcomb-type problems as the amount of money in B1 is varied (41-45).

The Prisoner's Dilemma

The Prisoner's Dilemma (PD) is a simple game in which two symmetrically situated players, ROW and COLUMN, must simultaneously and independently choose to cooperate or defect. Moreover, they both know the (typical) payoffs for their choices as outlined in the choice matrix below (listed as ROW's payoff, COLUMN's payoff):

	COLUMN Cooperates	COLUMN Defects
ROW Cooperates	1, 1	-2, 2
ROW Defects	2, -2	-1, -1

The Prisoner's Dilemma

Table 1

A player thus has the following rational preference ordering: (1) her own unilateral defection, (2) mutual cooperation, (3) mutual defection, and (4) her own unilateral cooperation. The PD thus models, in a simple and theoretically tractable way, many common instances of social interaction where two (or more) individuals have good reason to hope for both the other's cooperation and strong temptation to defect (i.e., industrial pollution, marital fidelity, taxpaying, theft, political attack advertising, etc.) (Heap 100, 144; Gauthier, *passim* ; Dixit and Skeath, 274-80).

But how should an individual make a decision when faced with a PD type

choice? Consider the situation for ROW. If COLUMN cooperates, she is rational to defect since this choice has a higher utility. On the other hand, if COLUMN defects, ROW's best choice is still to defect. So defection is ROW's dominant strategy. And since the game is perfectly symmetrical, the same is true of COLUMN. Therefore, in a one-shot game, dominance dictates that both should defect and thereby find themselves in the unenviable position described in the lower right hand box of the matrix.

However, in an iterated Prisoner's Dilemma, the situation is different, since each player has a reason to hope that the other player will cooperate in future rounds. Many strategies have been suggested to show why and how the players can move from the lower right hand box to the upper left hand box, but the simplest and perhaps most effective is tit-for-tat, which was first suggested by Anatol Rapoport (Davis, 148): Cooperate on the first round and then copy the other player's previous choice on the next round. Players who adopt this strategy will, in most cases, fare better than those who continually defect or continually cooperate (Heaps et al, 123-4; Dixit and Skeath, 259ff).

Even though NP and the PD are importantly different, they are, Nozick argues, sufficiently structurally similar to allow us to apply the same treatment to both problems. In each case, the dominant strategy that the causal theorist adopts will lead to a sub-optimal outcome, whereas, he claims, the evidentialist may fare better. That is, if one PD player thinks the other player is like her, then her own disposition to defect or cooperate is evidence of the other's similar disposition, and she can therefore predict (with a degree of certainty proportional to her degree of certainty that she and the other player are rationally similar) that the other agent will make the same choice she does. Knowing this allows her to eliminate the upper right hand and lower left hand boxes of the matrix, and the

resulting choice - between mutual defection and mutual cooperation - is a forced one. The causal theorist, of course, is more interested in the differences between the relative payoffs - overall - of defection and cooperation strategies (Rationality, 52-4).

But, again, Nozick argues that the formal nature of the problem that allows both the causal and evidential partisans to find support for their interpretations is problematized when we assign specific values to the payoffs. For example, in PDs where unilateral defection pays off only slightly better than mutual cooperation, and mutual cooperation pays off far better than mutual defection, parties will be more attracted to the evidentialist argument (see Table 2, adapted from Rationality, 53):[14]

	COLUMN Cooperates	COLUMN Defects
ROW Cooperates	1000, 1000	0, 1001
ROW Defects	1001, 0	1, 1

The PD - Incentive to Cooperate

Table 2

But if there is a huge gap between the payoffs between unilateral defection and unilateral cooperation, and where mutual defection carries only minimally less utility than mutual cooperation, parties will lean to the dominant choice and to

[14] Notice that cases in which payoffs for mutual cooperation are sufficiently high, or very close to the payoff for unilateral defection, also support a satisficing response. This should not blind us to the fact that the Prisoner's Dilemma is meant to clearly illustrate the problems of interaction between two players who do wish to maximize. If we lower the payoffs for cooperation (relative to those for defection) enough, even satisficers will be drawn into the same conundrum.

mutual defection (51-58). Table 3 (adapted from Rationality, 53) demonstrates this:

	COLUMN Cooperates	COLUMN Defects
ROW Cooperates	3, 3	-200, 500
ROW Defects	500, -200	2, 2

The PD - Incentive to Defect

Table 3

As in Newcomb's Problem, neither the causalist nor the evidentialist can rightfully claim complete confidence across the entire spectrum of PD cases, even where they are structurally identical. Therefore, Nozick contends, decision value should allow us to weight causal and evidential utilities as the case demands (53-4).

But has Nozick offered a credible defense of evidential utility? Nozick seems to suggest that if we lived in a world full of NPs, it would be rational (even for the causalist) to drop a pill that would make one into a CEU maximizer. David Christensen, however, warns us that there is an important distinction between the rationality of adopting a decision rule (or popping a pill that induces one to follow that rule) and the rationality of the decisions made under that rule. Imagine then a world in which the Thought Police routinely tortured individuals who took steps to discount their biases (as many accounts of rationality suggest we ought to do). In such a world, it would then be rational to not compensate for one's biases. But one's decisions made in this way, Christensen suggests, need not be rational. Likewise, in a world of many NPs where a predictor punishes

CEU maximizers, it may indeed be rational to swallow the EEU pill. But the benefits that accrue to the EEU maximizer are collateral benefits in that they are independent of the agent's decisions, just as, in the earlier example, the agent's avoidance of torture through making biased decisions is a collateral benefit of those biased decisions (262-3). The contents of her beliefs themselves do not benefit her (by helping her make rational decisions, for example), as is the case for responsible epistemic agents who seek to hold true, unbiased, beliefs.

But it seems to me that there is an important asymmetry in these examples between the rewards for bias and the rewards for evidentialism in these examples. In the first case, the agent is rational to avoid torture, but her belief-forming procedures are still biased and therefore may cause her to increase her degree of belief in some claim where an increase in belief is not warranted, and thereby to make decisions that will not maximize utility.[15] In sum, the Thought Police can punish unbiased decision makers but they cannot thereby change the world to make it one in which biased beliefs become true or in which biased decisions to act will reliably achieve their aims. The NP is different in that a disposition to EEU does change the world by (retroactively?) influencing the predictor's decisions to place the $M in B1. Hence one's decision to take the EEU pill and one's decisions under the influence of the EEU pill are both rational. It seems to me that a more felicitous analogy would be to suggest that one is rational to drop a cooperator pill in a world in which other agents will only cooperate with cooperators (see Gauthier Morals by Agreement, 169ff)) . In this case, both the decision to take the cooperator pill and the subsequent decision to cooperate are equally rational.

[15] See the end of Chapter VI for extended discussion of this point.

As Megone tells it, Nozick's case for EEU seems to rely heavily on the supposed existence of Electionist thinking among European Christians a couple of centuries ago and the even more dubious assumption that such thinking is rational (366). But, as my exposition of Nozick shows, Nozick can and does offer plausible reasons to consider EEU that are independent of any historical claims about instances of Electionism (even though these putative beliefs may have played a suggestive role in Nozick's motivations for defending EEU (46)).

There is in fact another reason to think that Nozick does not (and cannot) depend on Electionist reasoning to support evidentialism here. Susan Hurley has argued that both critics and defenders of evidentialist reasoning agree that if Nozick's account is to stand, there must be some difference between the legitimate appeal to EEU in NP and the PD, on the one hand, and the non-legitimate use of it in the Electionist and "career gene" examples on the other. The chief point of contention, she says, is whether the onus is entirely on the evidentialist to explain why they are different, or whether the causalist has also to explain why differing payoffs will alter most people's choices in NP/PD problems, but not in the latter examples. But this latter concern is also a problem for Nozick, Hurley argues: the evidentialist line of reasoning Nozick develops for the NP suggests that if evidentialist considerations are not always given zero weightings then a parallel phenomenon - shifting intuitions - ought also to occur in the Electionist and career gene cases (i.e., varying the utilities of material acquisition or of pursuing a genetically favored career ought to shift causal and evidentialist weightings) - but they do not ("New Take" 67-8).

Hurley's proposed solution is that it is cooperative, and not evidential, reasoning that explains the differences in the NP and PD. Agents are rational if they engage in cooperative action only if they understand that their collective causal powers

can bring about some jointly preferred result. In the PD, acting cooperatively will yield both players their second best outcome. However, in the NP, there is only an illusion of cooperation. Hurley thinks that choosers are implicitly reasoning somewhat as follows: "The predictor wishes most of all to not give me any money and that I not act greedily (by choosing both boxes) and wishes least of all that he pays out $M + $K and that I choose greedily. I, on the other hand, prefer (in descending order): $M + $K, $M, $K, and nothing. Therefore, by acting cooperatively, the predictor and I can both achieve our second best choice - that I act non-greedily and receive $M." In sum, the chooser has transformed Newcomb's Problem into a PD, even though the predictor's supposed interest in cooperation is illusory, since the NP specifies no preferences for the predictor. In sharp contrast, there is no room for any cooperation, or even the illusion of cooperation, in the genetic example, and this explains both the shifting intuitions over a range of NPs, and the disanalogy with the genetic example.

However, there are several problems with Hurley's analysis. The first is that while her account neatly explains the divergence of our intuitions between the NP and career gene examples, it does not so clearly explain why we think evidentialist reasoning is permissible in NPs but not (as Hurley thinks) in Electionist cases. After all, the Electionist can argue that both causal and evidential considerations concur in pursuing materialism, since the choice to do so dominates material deprivation whether or not material wealth is associated with heavenly rewards. So if cooperative reasoning is plausible within the NP, why not for the Electionist?[16]

[16] Can a satisficer avoid these conundrums? Minimally, any rational Electionist will wish at least to avoid eternal Hell. Since the only way to avoid Hell is to ensure that one enters Heaven (which is maximally good), the rational satisficer should therefore act exactly as the maximizer does.

Second, Hurley does not explain why, if the NP provides only the "illusion" of predictor cooperation, players who rely on this cooperation will fare better than causalists. If the predictor isn't in fact cooperating, what features of the problem explain the evidentialist's success - and couldn't those features be the relevant ones that also explain her decision? Moreover (and this may be the answer to my last question), Hurley fails to appreciate the way in which cooperative reasoning itself relies on evidentialist reasoning. To the degree that a player is disposed to cooperate, and knows that the other thinks as she does, it is to this degree that she can rely on her own disposition to cooperate as evidence of the other's willingness to cooperate. Political theorists have tended to emphasize the role of detection and punishment of defectors as a way to influence decisions in the PD. If agents believe that there is a sufficiently strong probability of a sufficiently undesirable outcome for their defection, they will be more likely to cooperate. But this fact alone does not explain why agents frequently cooperate without regard for penalties (because they have internalized the disposition to cooperate). Presumably, part of the reason why an agent might be willing to conform to a given set of behavioral expectations is that she believes that most other agents are also disposed to conform to the same expectations, and that she also believes her disposition to do thus and such is therefore evidence that others will do likewise. It is even plausible that evidential reasoning from an assumption of conformity plays a larger role in securing widespread cooperation than the purely causal effects of public sanctions. And this is so even where both players have strong reason to cooperate and no reason to defect.

Imagine, for example, two individuals each hoping to find the other at a large venue (a university campus, large shopping mall, or outdoor exhibit). If one or both of them have no idea of the other's likely behavior, they will find each other only by chance. On the other hand, if they both reliably behave in the same way

in this sort of situation, and each knows that the other is likely to behave in this way, and each knows that the other knows, and so on, they then have a much greater chance of locating each other: "I would wait by the fountain, so that's probably what she would do. And since my disposition to wait by the fountain is evidence of her disposition to likewise do so, my willingness to do so counts also as a good reason to do so." So, we might conclude, Hurley has not provided a convincing argument to show that evidential reasoning is in any way otiose.

David Christensen, Paul Moser, and Christopher Megone have all argued that Nozick faces deeper methodological problems in his approach to rationality. Christensen wonders why common intuitions should carry any weight, given the plethora of studies showing that many people - including supposedly sophisticated professional reasoners such as philosophers and mathematicians - routinely make obvious errors of logic and induction (261-2). And Megone observes that even though the stated object of Nozick's study is the nature of rationality, Nozick himself never seems to give us a clear account of the means by which we might determine what counts as rational or otherwise. Although Nozick (like Aristotle) counts rationality as a distinctively human capacity, and therefore relies on observations of human behavior as one way of giving shape to his concept of rationality, he also discounts some instances of very widely spread behavior (such as succumbing to temptation) as irrational. By what standard, Megone asks, can Nozick then build a case for some aspect of rationality by appealing to its widespread acceptance in one case, while denying that some behavior is rational even though it is widespread (363-5)? Megone is equally critical of Nozick's use of the NP and the PD as ways of explicating a normative theory of rationality. Insofar as these problems do not model real-life situations, he argues, the ways in which agents should resolve them are hardly pertinent to rationality (369-70).

But Nozick offers at least a partial answer to these worries. "In order to justify a principle, you specify its function and show that it effectively performs that function and does this more effectively than others would given the costs, constraints, and so forth (36)." Such a justification must allow that the function is desirable, and although Moser wonders how it can be shown in any normative sense that a function is "desirable", it seems clear from Nozick's heavily instrumental account of rationality that a function's desirability is to be determined by the agent's own goals and plans: if she favors certain goals, she will favor principles that function so as to bring those goals to fruition. And instrumental rationality, claims Nozick, is the default position of all theories of rationality: whether or not a theory can successfully defend any expanded (non-instrumental) view of rationality, it must at least recognize rationality's instrumental role (133). But it may be that the principles that we propose as a way of managing choices (in the way that scientists discover principles as a way of managing scientific prediction) may not, in the long run, prove as effective at maximizing utility across a very large and disparate set of interests as the problem-solving heuristics which natural selection has favored. As is familiar to students of biological design, the blind and ruthless forces of nature are often far more ingenious and effective - in quite unexpectable ways - than the highly focussed attention of intelligent human designers. So here we can find room for the legitimacy of Nozick's appeal to evolutionary forces.

The second part of the answer lies in the suggestion that there is a difference between providing a defense of rationality itself and a defense of an account of rationality. As I read Nozick, he is engaged on the latter, and not the former, project. Defending a concept of rationality requires only that the concept is itself rationally acceptable. Where the opinions of individuals differ, then - insofar as

they are rational - argumentation can lead them to converge on a common understanding. Thus one can appeal to shared agreement in some cases while still affirming that many people do, in other cases, act irrationally if one believes that individuals would, in those latter cases, revise their judgments given enough time and rational persuasion. And in cases of temptation, one can argue that one's dispositions to resist temptation at T1 and T3 do in fact count as rational, while one's disposition to succumb at T2 does not, simply because one has a second-order desire at T1 and T3 to not have a desire to succumb at T2, while one does not have a similar second-order desire at T2 to efface one's resistance to temptation at T1 and T3. So Nozick's appeal to shared opinion may not be as problematic as critics take it to be. And yet, despite all this, there may be no unanimity on some important questions, and no decisive method whereby all parties could agree that a decision could be reached. In these cases, where there is no test of what counts as rational beyond what rational thinkers consider to be rational, and where they do not agree, it may be that the only defensible position is to strike a compromise which gives due weight to equally compelling intuitions on both sides. And this, I think, is what Nozick has attempted to do in the theory of decision value.

Symbolic Utility

Nozick's other chief emendation to decision theory (and the principle focus of this work) is his suggestion that decision value ought also to recognize the symbolic utility (SU) of an action. On a theory of decision theory which does not recognize symbolic action, an agent undertakes the action she does (and attributes some value to it) because the utility of the goal which that action is to bring about is imputed back to the action. So,(as in Nozick's example) our agent might be able to overcome temptation in a given circumstance because she recognizes that to do so once may play a (slight?) causal role in being able to overcome temptation in the future. Or her resisting now may be evidence that she can resist temptation later, and that evidence itself becomes a reason to forego the temptation now. These both count as examples of what I have called narrow rationality. And, yet again, resisting temptation now may symbolize - in a way that is distinct from the causal or evidential roles that act may play - the value of future acts of temptation-resistance. Accordingly, symbolic utility's (SU) value must be noted alongside the causal and evidential values. Our decision value formula thus becomes:

5. $DV(A) = Wc \times CEU(A) + We \times EEU(A) + Ws \times SU(A)$

Nozick is emphatic that symbolic utility is not a sort of utility which differs in kind from causal or evidential utility, since an action taken for symbolic reasons derives its utility from the same sort of utility (48). It is only the connection which differs. Christensen, however, has argued forcefully that there is no good reason to make any revision to decision theory to accommodate symbolic utility as a different sort of utility, and we should attend to this objection before proceeding with a more detailed examination of symbolic utility.

As Wes Cooper and I recently pointed out in response to Christensen, "this [unwillingness to consider a separate "bookkeeping" system for SU] is a formula for inconclusive rounds of shifting the burden of proof" and we likened Christensen's attitude to to early American computer programmers who resisted second generation programming languages (which greatly facilitated programming by translating machine code into a more "natural" language) simply on the grounds that there was no formal reason to do so (Cooper & Barron, 12). And since Christensen offers no clear criteria for what would count as a positive reason to differentiate SU, it is not clear that all of his objections are as compelling as he thinks. Nonetheless, they merit our attention. In brief, Christensen finds three arguments in Nozick to defend the symbolic /nonsymbolic distinction: (1) we may wish to track the varying weights of SU in different choice situations (48), (2) the SU of a goal may not vary proportionately to the probability of achieving that goal (34) and (3) the SU of a given act may be influenced by what other acts are available to the agent and the utilities of those acts (55).[17] A final point (which Christensen takes to support (3) but which appears to be a separate concern) Nozick expresses thus:

> Many writers assume that anything can formally be built into the consequences, for instance, how it feels to perform the action, the fact that you have done it, or the fact that it falls under particular deontological principles. But if the reasons for doing an act A affect its utility, then attempting to build this utility of A into its consequences will thereby alter that act and change the reasons for doing it; but the utility of that altered action will depend upon the

[17] So, for example, the symbolic utility of Socrates' acceptance of the death penalty might be seen to vary according to the alternatives available to him and to their respective utilities.

> reasons for doing it, and attempting to build this into its consequences will, alter the reasons for doing that doubly altered act, and so forth. Moreover, the utilities of an outcome can change if the action is done for certain reasons. What we want the utilities of the outcomes to represent, therefore, is the conditional utilities of the outcomes given that the action is done for certain reasons (55).

Against (1), Christensen simply observes that other forms of utilities also vary contextually and that Nozick has therefore failed to point out any meaningful difference between SU and other types of utility. Christensen similarly replies to (3) by saying that other forms of utility may also vary when regarded against the backdrop of the agent's available options - what counts as courage, morality, or rudeness in any given situation may depend on what other choices the agent had available to her. Similarly, even an agent's nonsymbolic reasons for performing some action may affect its utility. If, for instance, she attends a concert merely to be seen, or if she has sex for money, she may be less likely to enjoy the aesthetic or erotic pleasures of these activities (265-7). Given this, the fact that reasons for undertaking symbolic actions affect their symbolic utility does not distinguish them from nonsymbolic actions.

Consider now Christensen's response to (2). Christensen points out that if the utility of performing a symbolic act (say, defending the Alamo) is independent of actually realizing the state of affairs symbolized by that action (presumably, in this case, keeping the Alamo in American hands), this suggests that the utility does not in fact flow back from the state of affairs but arises merely from the value of symbolizing itself. (I consider this metaphysical question more

completely in chapter V.) In any event, he says, it would appear that the symbolic utility of the act is constant no matter what its outcomes are. "[W]e have seen no reason to think that there is any problem with taking all of an act's possible outcomes to include the same symbolic utility (265)." But to admit this much is, I think, to substantially concede Nozick's point. If the symbolic utility remains constant across all possible outcomes (winning or losing the battle of the Alamo), while other utilities obtain in only one outcome then SU cannot fall under the scope of the probability functions which affect the decision value for those other utility functions, but not for symbolic utility.

Still, some forms of symbolic utility are outcome-dependent (the symbolic utility of writing a dissertation which symbolizes my own self-overcoming is realized iff I complete the dissertation), while others (like the red roses Canadians placed almost everywhere to mourn Pierre Trudeau's death) are attached to actions that do not seem to be directed at any other nonsymbolic purpose which may succeed or fail. Nonetheless, the deeper metaphysical point about symbolic connection is this: while agents may fail to accomplish some symbolic action (say, by not buying a rose to mourn Pierre Trudeau), or may fail to understand completely the symbolic import of some action for other people, or may fail to fully communicate some symbolic value to another, there is a sense in which at least some symbolic actions, just in virtue of being symbolic, cannot fail to achieve their goal. That is, while causal actions typically derive their value from bringing about some other state of affairs, symbolic actions need only to achieve their own performance.

Despite all this, the most telling objection against a separate bookkeeping system for symbolic utility is that conventional decision theory treats an agent's decision in a behavioristic and instrumental fashion. That is, decision theorists are

agnostic about how exactly agents combine utilities. On this view, there is simply no fact of the matter except how the agent in fact acts and therefore all utilities must be folded into the consequences of an action, and these are completely describable in an agent's preference orderings. Utilities, then, are not "in the head," they are simply comparative measures employed to express how an agent chooses. To say that there is a separate "symbolic" component to an action (say, preferring a honest ten dollars over a stolen ten dollars) makes psychological claims which many decision theorists say are insupportable. It follows from this that the only way to quantify symbolic utility is by asking agents how much utility of another sort they would be willing to swap for the utility of some purely symbolic action.

Let us consider the nature of symbolic utility from another perspective. For an agent to decide to take an action on symbolic grounds, the action must represent or mean (or evoke) something beyond itself (which Nozick calls "M"), and that M has some utility for her. The fact that an action represents M will not, by itself, be sufficient to move the agent to action unless M has this utility and M's utility outweighs the utility of not performing the action (26-7). As I've noted earlier, decision theorists typically ascribe an outcome's utility for a given actor in a behavioristic fashion: they do not assume that utilities are things in the heads, but instead describe the actor acting as if she assigned such-and-such a utility to a given outcome. It is not clear that Nozick's claim that acts have a distinctive symbolic connection between them and an actor can be described without remainder in a behaviorist manner. Insofar as Nozick invokes Freud's account of neurosis as an example of a (unhealthy) class of symbolic actions (26-7), he seems committed to positing real mental entities as causally necessary in forging a symbolic connection between an agent and her symbolic acts.

Nozick also points out that symbolic actions are frequently expressive, and he uses this observation to construct a second, but parallel, understanding of the relationship between a symbolic action and what it symbolizes. On this second account, what flows back to the action is not symbolic utility but expressiveness (among other things, 28, 186).[18] The important difference between these two ways of understanding symbolic action is that the strength of the symbolic connection between an action and what it represents (presumably) remains constant while its expressive value does not, since it varies with the agent's moods and dispositions over time. For example, hand washing may symbolize or represent (to the neurotic) washing oneself free of guilt. Moreover hand washing always symbolizes guiltlessness and presumably being guilt-free always has a high utility. Since the symbolic connection is constant, hand washing should impute the same utility back to the actor. And yet the neurotic may not always wish to wash to wash his hands. This, argues Nozick, is because recent hand washings make the problem less acute, since the utility of expressiveness varies from context to context, and recent expressive hand washings may reduce the need for ones in the near future while other competing utilities are being sought (28). If the only connection between an action and what it represents were one of utility, we could not explain why people only occasionally perform that action (since its utility would be constant).

But this seems to me a needlessly cumbersome way to explain what is in fact a common aspect of utility. The utility of food, for example, is stable so long as its flavor, caloric, and other nutritional values remain constant. But even though

[18] A symbolic action may convey meaning by both representing something and by expressing it - both modes are communicative. Nelson Goodman says expressions differs from instances of reference in that they are of feelings or other properties rather than events or objects. Expression is also less literal, and is shown by intimation rather than imitation. Symbolic (especially artistic) expressions express by being exemplifications of what is expressed (Goodman Languages of Art, 45-52).

food's utility may remain constant, we do not eat constantly, because our appetites vary with the amount of food consumed. Another way of putting it is to say that food has both an objective utility based on its measured nutritional properties and a subjective utility measured by the the agent's tastes and degree of satiety. In the same way, one's "appetite" for symbolic expressiveness may vary over time, allowing us to understand why we do not constantly engage in symbolic action without needing to posit any two-track symbolic/expressive link between an act and its meaning.

For Nozick, the importance of symbolic utility bespeaks itself most clearly in its role in overcoming temptation. Nozick argues that creating normative principles for oneself allows one to attach symbolic weight to particular actions, allowing that one action to stand for many others. That is, if I wish to avoid eating sweets now, but know that I will be tempted to do so at some point in the future, I can adopt a principle linking these actions together. Nozick thinks that to adopt a principle such as "never eat snacks between meals" allows one act to stand for the rest. "[I]t is as if you have made the following true: if you do this one particular action in the class, you will do them all (17)." Prior to adopting this principle, doing the act did not necessitate doing the rest, but once you have accepted the principle, then to do the act "means [you] will continue to do it in the future (19)." It is perhaps a rather ambiguous infelicity on Nozick's part to say that performing the act "means" one will perform the others, and it does not help to say that formulating a principle covering these acts as a class "is a way of "tying" their consequences together (19) or that one action "speaks of" the others (21). What can all this mean? Nozick suggests that where one has not formulated a principle, performing one act may increase the probability of repeating it though what he calls the psychological "law of effect", whereby the positive reinforcement subsequent to that act increases the likelihood of its repetition. But

recognizing that the act and its possible successors are united under a single principle establishes a different sort of connection between them, such that performing the act or not will alter the agent's estimate of the likelihood of her repeating the action (19-20). And the combined disutility of all those actions may count as a sufficient reason to not succumb to temptation now.

However this way of expressing it is still ambiguous between saying that performing this act now counts as evidence that I shall repeat it and saying that doing this act now symbolizes repeating it. I don't think Nozick is particularly explicit on this point, but one way of making the distinction is to suggest that the evidentialist will avoid the first act because the principle she has adopted or recognized makes it obvious that if she undertakes the first act she is more likely to commit the others, and she wishes to avoid their disutility. In contrast, to posit a symbolic connection between these acts may be to say that the agent wishes to avoid the disutility of violating the principle itself: independently of the disutility of the projected future acts, she does not wish to be the sort of person who violates this particular principle under which those acts are grouped:

> This, I think, is what enables principles to define a person. "These are the lines I have drawn." It is these lines that limn/delineate him. They are his outer boundaries (Nozick, 26, emphasis in original).

Of course, if overcoming temptation were as easy as this, those struggling with addictions would have discovered this stratagem long ago, and this suggests that it cannot be quite as easy as Nozick suggests. To be fair, he does recognize that we cannot force upon ourselves any principles which tie together any actions, and that if we try to convince ourselves that any failure on our own part in any

situation whatever can in itself symbolize any future failing, we may be completely demoralized by a weakness of will, and we are therefore wiser to restrict our efforts at overcoming temptation by using symbolic utility to those cases where we are more likely to succeed (19-20). The question, however, is whether one can ever overcome temptation in this way. The problem here is perhaps that it seems psychologically improbable to think that one can simply decide, by mere fiat, to believe in a principle which unites the utility of one act with the utility of many other similar ones. In any event, as I argue in chapter III, intrapersonal applications of symbolic utility don't seem to count as paradigmatic cases of symbolic action and are therefore largely irrelevant to explaining why we perform them.

Similar considerations also explain why symbolic utility can attach itself to moral actions. An agent's moral action not only does what is good or right, it also symbolizes her commitment to a class of similar action and places her on the side of the values which express those actions (29-30). If Immanuel Kant is right, and moral principles count for more than their teleological value, our commitment to moral principles independent of their consequential value may symbolize our commitment to rationality itself (40).

What We Still Need to Know About Symbolic Utility

Nozick's account of symbolic utility, although suggestive, is hardly complete, and this much he fully recognizes. As Paul Moser asks:

> What kind of (psychological or sociological) mechanism yields symbolization of the kind prized by Nozick? Second, what precisely is the

> relation of A's symbolizing X that Nozick has in mind? To the extent that these questions remain unanswered, talk of symbolic meaning and symbolic utility is obscure. (288)

Moser's first question can be taken to require an explanation of two phenomena: how symbolic utility arises from individual psychology and the process by which it then becomes part of a social fabric and how and why individuals receive this meaning into their own individual lives. Moser goes on to suggest that in some cases what Nozick describes as SU may be nothing more than the expression of second-order desires and therefore can be more properly understood within the usual nexus of causal utility.[19] In some cases, this may be true, and Nozick's example of overcoming the temptation to smoke seems a case in which Moser's point appears especially telling: it is more plausible to reinterpret such cases by saying that the individual has a second-corder desire at T1 not to have first-order desire to smoke at T2. But many symbolic actions (attending funerals, for example) do not seem aimed at overcoming first-order desires.

And in the absence of any detailed account as to how symbolic links are forged, Moser wonders if such links could be forged between any two arbitrarily chosen objects (289, 291). We have also to distinguish communicative symbolic action (such as language) from other forms of symbolism which derive their utility from their non-communicative role. That is, speech acts (usually) possess causal utility in virtue of their power to convey information to other individuals, and this causal power is a measure of their communicative utility. But speech acts frequently have symbolic utility (where we understand "symbolic" in the way

[19] See Harry Frankfurt's The Importance of What We Care About for discussion of second-order desires.

Nozick has defined it) that may be independent of communicating anything to any particular individuals. That is, we may have no audience in mind when we perform some symbolic speech act, or the audience may not learn nothing new (directly) from the content of the speech act. We need to understand also why symbolic action of this sort is universal to the human experience, and why it might have roots in human evolution. A further consideration is the ontological status of symbolic actions within a human social reality, how symbolic actions can be construction from and derived their distinctive status from human intentionality, and how we define their function.

CHAPTER III

The Evolution of Symbolic Utility

Simply the thing I am shall make me live.[20]

If the line of argumentation in Chapter VI (below) is correct, then it is plausible to think that the mind is a collection of domain-specific modules which are as they are in large part because they were selected for their contribution to fitness in the ancestral environment. Moreover, this hypothesis can be used as a heuristic to reverse-engineer the likely contribution to fitness of many particular aspects of rationality - such as the capacity to recognize and respond to symbolic utility - as a way of understanding the ontogeny of that capacity. Since symbolic action is uncontroversially a part of all human societies and since it has deep and strong connections to central and universal human concerns, it is more than plausible to suggest that evolution has endowed humans with a specific capacity for symbolic action. So my ambition in this chapter is to augment Nozick's account of the evolutionary roots of human rationality by suggesting why symbolic utility is both possible and prevalent in human action. And thereby to explicate how symbolic meaning moves downward from the social to the personal level. This will entail (1) delineating just what counts as a symbolic action, (2) specifying the selective forces that played a role in the creation of the human symbolic capacity, and (3) suggesting why a capacity for symbolic action may have contributed to fitness.

[20] Shakespeare, All's Well that Ends Well. (1603-4). act 3, sc. 3.

The Nature of Symbolic Action

From chapter II, it will be obvious that while Nozick considers the utility derived from symbolic actions to be identical to the utility derived from other sorts of actions, he thinks that the connection between symbolic acts and their outcomes differs importantly from evidential or causal connections (Rationality, 48). To borrow Nozick's most illustrative example, my having this one snack, here and now, will not cause me to continue snacking. Nor is it necessarily evidence that I will continue to do so. But, "by adopting a principle," I can allow this one snack to stand for many other future snacks, and this snack thereby acquires the negative utility of all my continued snacking (18). So this occurrent act, in virtue of its being a token of a similar type of action, exemplifies that type of action in such a way that (as Nozick variously puts it) it represents those actions, or in such a way that the "meaning" of this act is those other acts, or so that expressiveness flows back from the outcomes thus symbolized (33). In fact, most of Nozick's examples display this token-type[21] structure wherein one act exemplifies some desired state of affairs which it also represents, although he is quick to add that the link may in some cases be less direct (33):

- Anti-drug laws and drug use reduction (27)
- Minimum wages laws and helping the poor (27)
- Saving an actual trapped miner and saving all actual victims (as opposed to saving "statistical" lives through allocating resources for accident prevention) (32)
- Believing one falsehood and believing all falsehoods (71)
- committing one moral act and being an ethical agent or a member of

[21] This distinction dates to C. S. Pierce, but is used much more broadly than he originally intended it. A token is a particular instance, of a given type of abstract object. So my act of typing here and now counts as a token of the type typing. See Jennifer Hornsby "Token," 877.

the kingdom of ends, etc. (29, 63)

- Repeated hand-washing and ridding oneself of guilt feelings (26)

It seems from this (though Nozick is not explicit on the point), that one might reduce linguistic utility to causal utility in this way: the causal utility of my uttering some sign for *p* will be determined by the utility for me of making the listener believe *p*, and the causal utility of hearing some intentional sign for *p* will be determined by its evidential value in causing me to believe *p* (49-50).

In a very similar fashion, Nozick adopts Nelson Goodman's account to explain the various ways a symbol may be connected to its referent:

> *A denotes B* when *A* refers to *B*; *A exemplifies P* when *A* refers to *P* and *A* is an instance of *P*, that is denoted by *P* (either literally or metaphorically); *A expresses P* when *A* refers to *P* figuratively or metaphorically (so that P figuratively denotes A), and in exemplifying P, A functions as an aesthetic symbol (33).

And these connections can be linked together to form chains of symbolic connection. It seems to follow from this that M - the property which is imputed back to an agent via a symbolic connection -and which creates symbolic utility itself must inhere in the <u>difference</u> between linguistic and symbolic connections. This is so, it appears, because there is nothing in linguistic symbolism, construed as <u>purely</u> communicative behavior, which could confer some benefit *M* on a speaker, over and above the utility which the speaker derives from changing the beliefs, behavior, desires, etc., of a listener. But communicative utility is a purely causal utility and, as such, the speaker should be rationally indifferent to which modality she employs to effect changes in the listener's mental states, and the

utility of her speech acts is exhausted by the utility of changing the listener's mental states. Symbolic actions, on the other hand, do derive at least part of their utility from the ways in which they are performed, and need not depend on either anyone observing them or on any changes to the spectators' mental states.

Broadly speaking, of course, all linguistic behavior is symbolic, since language is perhaps the most highly developed symbolic behavior which humans possess. But Nozick -and I - interpret "symbolic" more narrowly. Plausibly, what Nozick intends by symbolic utility is all that utility which relies on a "meaningful" or "referential" link (as described above) over and above any linguistic-causal utility it may possess. Pretty clearly, not all symbolic actions as defined in this way are communicative. If they are undertaken privately, they have no audience. And in many shared and public symbolic actions, the symbolic utility may not be dependent on making the audience come to believe *p* (since they may already believe *p*, or they disbelieve *p*, but the actor does not care to convince them of *p*). In other cases, an action may have symbolic utility only in virtue of its also communicating some claim to another. For example, uttering a marriage vow only carries a symbolic value if a certain audience hears it. And telling your employer, "I quit," may have a self-asserting power that is greater than the merely communicative utility of informing the employer that you will not be coming in tomorrow.

What is obscure in this account is exactly why symbolic connections transmit M (for "mystery"?), why purely linguistic ones do not transmit M, and why humans should feel a need to engage in symbolic behavior at all. I hope the answer will become clearer later in this chapter. (This is not to imply that some speech acts - such as marriage vows - cannot also have symbolic utility. I am merely making

the point that symbolic utility is not necessarily intrinsic to all speech acts.)

After all, as Nozick remarks, the fact that agents appear to be acting in ways that are unlikely to maximize garden-variety causal utility is a strong indication that their actions are symbolic (27).[22] And in many cases, pursuing symbolic utility may decrease the causal utility which it represents: a soldier who dies to protect his nation's flag (because it represents national ideals which he believes in) thereby destroys his ability to defend those ideals themselves. So, given the universality of symbolic action, its obvious importance in most people's lives, and the sizable losses of other utilities which it sometimes occasions, there must be some deep and widely influential cause or causes why we act symbolically. Nozick thinks that we pursue symbolic utility because the very capacity for symbolization is one we hold dear:

> A large part of our lives consists in symbolic meanings and their expression, the symbolic meanings our culture attributes to things or the ones we ourselves bestow. It is unclear, in any case, what it would be like to live without symbolic meanings, to have no part of the magnitude of our desires depend upon such meanings. What then would we desire? Simply material comfort, physical security, and sensual pleasure? And would not part of how much we desired these be due to the way they might symbolize maternal love and caring?[23] (30)

I don't think that Nozick, in making these comments, abandons his strong

[22] Of course, it is also possible that their actions are simply irrational.

[23] Does Nozick intend to include linguistic meaning within these "symbolic meaning"? I think not, since Nozick stresses that symbolic meanings need not all be "good" ones (in the way that desires or preferences might be good or not good) - and this is not a predicate which one would normally bestow on a purely linguistic meaning.

commitment to instrumental rationality. To perform an action for symbolic reasons is not simply to do the act for the sake of doing it. On the nonsymbolic account of eating, for example, we eat because we find the taste of food pleasing and/or to satiate feelings of hunger. But suppose one eats for symbolic reasons, perhaps in a religious ceremony. Does this mean her actions are thereby not aimed at some end? I think the answer is no - the fact that her motivations are not the usual ones does not mean they are absent. The fact that she may not be clearly aware of her desired goal does not mean it does not exist. (After all, we do not have a vocabulary which describes the senses of well-being or unease that accompany the completion or noncompletion of symbolic actions as we do for the feelings that attend the completion or noncompletion of nonsymbolic actions - hunger, satiety, fear, feeling safe, thirst, loneliness, etc.)

Of course, the fact that the nonexistence of some psychological feature (such as the disposition to perform symbolic actions) is - any sense - unimaginable does not explain why it should exist at all. It is not a necessary fact of human existence that we or any other rational social beings should perform symbolic actions. It is in fact quite contingent on our specific biological and social conditions. So Nozick's comments do not explain why we should find it inconceivable that we not undertake symbolic action, and offer no help in explaining its existence. Nozick offer a slightly different view later:

> If human beings are Humean beings, that seems to diminish our stature. Man is the only animal not content to be simply an animal.[24] (Since my argument is motivated, you - and I too - should be alert to correct for any biases in its treatment of reasons.) It is symbolically important to us that not all of our activities are aimed at satisfying our given desires .... One

[24] See the Shakespearian quotation at the beginning of chapter V.

> way we are not simply instrumentally rational is in caring about symbolic meanings, apart from what they cause or produce. The proponent of instrumental rationality cannot easily claim that such caring is irrational, for he has no relevant criterion of rationality - why then should this caring be any more irrational than any other? Symbolic meanings are a way of rising above the usual causal nexus of desires and it is symbolically important to us that we do this. (138-9)

The idea here is that symbolic action is itself self-subsuming (139). To paraphrase Nozick slightly[25], suppose there is some theory *T* that describes the properties of any rule that defines rational action. And suppose *T* itself has those properties. Then it would be the case that *T* would validate itself by subsuming itself under *T* (Explanations, 119-121, 131-132). Then, since engaging in symbolic action (rather than not) has the same self-expressive properties that validate performing particular symbolic actions, a theory validating some particular set of symbolic actions will also validate symbolic action in general. But, as Nozick admits, self-subsumption is "quite weird (Explanations, 120)" and if we can possibly do so, it is perhaps better to avoid such, well, metaphysical justifications. So my intention is to provide a more naturalistic account of symbolic action that avoids such stratagems.

[25] Nozick employs self-subsumption to explain how very deep-level explanatory theories can validate themselves, thus avoiding infinite regresses. He says, but does not explain how, that the same process is available for symbolic action and instrumental rationality.

Sperber and the Non-Meaning of Symbols

So Nozick's account, as I read it, does not accord any necessary role to social groups in forming symbolic meaning, suggests no mechanism (other than simple volition) why symbolic connections should exist here and not there, and does not explain why the relatively minor structural differences between linguistic connections and symbolic connections are responsible for such radical changes in kinds of utility. In all this, I think Nozick would concur, and would welcome a richer account of symbolic utility. So what I want to do next is to offer an alternative account of symbolism and symbolic action largely borrowed from the anthropological work of Dan Sperber which will show that Nozick's account is not only incomplete, but unable to provide the explanatory framework that I think a complete understanding of symbolic utility requires. This, I think, will count as an important emendation to Nozick's theory of symbolic utility.

Nozick distinguishes between linguistic and symbolic utility, but only offers a partial analysis of the former. Following Paul Grice, he defines a natural (i.e., nonlinguistic) sign as one which is evidence for *p* (for example, dark clouds are a sign of rain, thunderclouds are a sign of rain, bared animal teeth are a sign of aggression,etc.) Signs acquire their relation with that which they signify because they regularly correlate with the signified. But a person could also intentionally produce some sign, a linguistic gesture, as a sign for *p*, and do so with the intent to make some other person believe *p*. (For example, a red octagon means "stop", a "?" indicates the interrogative mode, and so on.) This sort of intentional sign then correlates loosely with what C. S. Pierce counted as a "symbol", insofar as it is created for and understood as performing this denoting function (Pierce Laws of Logic, in Kolak). Both symbols and signs stand for or represent in some way some other thing, but the relationship between a sign and what it represents is a

"natural" one (C. S. Pierce calls the relation a "correspondence in fact" (Categories, sec. 15, in Kolak)). The relation between an symbol and what it signifies, on the other other hand, is an arbitrary one. The presence of a symbol need not always be accompanied by the presence of the signified. As Alison Jolly points out, symbols refer to what is "not here, not now, not present" and this property allows them to become at least partially detached from their referents ("Communication", 167, 175). Thus, since symbols, understood in this way, no longer have a direct causal connection with the phenomena which they represent (as signs do), but are really about the intentional states of their producers, the opportunity exists for deception. A second distinction within the use of symbols is that between what J. S. Mill called connotation and denotation (System of Logic, 19-25). (These two terms correspond very roughly with what Gottlob Frege called *sinn* ("sense") and *bedeutung* ("reference").) The denotation of a term is the class of things to which it it refers. So the term "man" denotes Socrates, Paul, Samson, etc. But the term "man" applies to them all equally because they share some set of qualities which the term "man" connotates.

Sperber's first point of departure from this account of symbols is to insist that we cannot construct a grammar for symbols in the way that we construct a grammar for languages. He offers four reasons why symbolic practices diverge importantly from linguistic ones. First, the stimuli that an individual interprets as linguistic data constitute a more or less homogeneous and distinct set of sensory inputs. Phonetic data are perceived audibly (or as written text) and organized in such a way that one is rarely uncertain as to what counts as a meaningful utterance. But this is not so for symbols, which are heterogeneous and diverse in their manifestations, without systematically common properties, and presented to us as myth, ritual, art, adornment, gesture, etc., indiscriminately through all

sensory modalities. There is thus no obvious perceptual way to discriminate between symbolic and nonsymbolic actions. And Sperber points out that often the anthropologist's most difficult task is therefore to delimit exactly what is a symbol and what is not. While some rituals may be clearly set apart from everyday life, many other symbolic practices are interwoven with everyday activities with no obvious (to the outsider) markers to set them apart. And where markers of the symbolic do exist, they too must be interpreted symbolically.

Second, language acquisition relies on a more or less fixed set of data to interpret in that one learns the language as it is spoken at a given time and place, and one does not typically use a sentence in Chinese, for example, as a datum in constructing an English grammar. Different children learning a common language will typically do so by hearing u sentences, but they will nonetheless converge on a common grammar. In contrast, symbolic data are more likely to be shared between many people, in that individuals will acquire symbolism by seeing and participating in the same rituals. But since symbols are not so clearly delimited from other facts of social life, and since there is no clear criterion for a given datum's inclusion or exclusion as a symbol, individuals will vary greatly in the ways that they process symbolic input and accordingly will <u>not</u> converge on a common understanding or a common symbolic grammar. It may not be possible to <u>measure</u> these differences, but they are evident to any observer in the common disputes over public symbols (i.e., the monarchy, etiquette, flag-burnings, to name but a few).

Third, since the grammar of one language cannot be used to interpret another language (unless, of course, they are closely related), learning a new language always entails the acquisition of a new grammar. But new symbolic data does not create a new symbolic mechanism: new symbolic data are always interpreted

by the same symbolic mechanism[26], which is itself modified only by the new symbolic data. This is, for example, why the anthropologist can switch with ease from one language to another (assuming he has learned both of them), but will find it harder to jettison his symbolic assumptions when moving from culture to culture. One cannot help but internalize the sense that certain acts are rude, polite, etc., and it is difficult, even in another culture, to reinterpret these acts in another way.

I can attest to this from personal experience. After four years of work in West Africa in the late seventies, I was completely fluent in Krio, and could speak passable Temne and Mandinka. However, I never achieved the same degree of cultural fluency. On my return to Alberta, I had no difficulty in speaking my native language. But I occasionally found myself reacting to social situations and gestures very much as a Sierra Leonean might have.

Fourth, linguistic grammars, once learned, are not modified by supplementary data, which only expand a person's linguistic skills. But, in contrast, new symbolic data continue to modify the symbolic mechanism, and there is no clear threshold at which one becomes "competent" in the way that one is linguistically competent (86-91). Maykel Verkuyten, who follows Sperber on many points, describes the disanalogies between language and symbolic action in a slightly different way as part of his study of the symbols associated with the Gulf War:

> First, symbols encompass different meanings (e.g., freedom, equality, national sovereignty) as a totality, as a connected whole that presents itself instantly and all at once.... Language has a discursive character

[26] Sperber is not clear about the exact nature of the symbolic mechanism. I am interpreting it as a psychological module of the sort which is described in Chapter VI in the section entitled "Heritability of Psychological Features."

> and its form requires that the thoughts succeed each other, even if that what is talked is indivisible. Symbols focus more on the totality....
>
> Second, symbols may fill in gaps in the lexicon, permitting people to experience and communicate what is beyond the bounds of existing speech: symbols 'lead us to realms of wordless thought' ... Symbols capture those realities that are not effectively expressed in all their 'thickness' by the conventional use of words.
>
> Third, effective social symbols do not only have a cognitive, but also an emotional meaning, as the examples of the burning of the American flag [to protest the U. S. involvement in the Gulf war] and the picture of the bird in oil [which many respondents said symbolized the innocent sufferers of the Gulf war] In the tradition of structural anthropology, cognitive anthropology, and semiotics, there is a clear emphasis on thinking and cognitive processes. Symbols are treated as a kind of arcane sign language that must be deciphered. The powerful emotional charge that most social symbols carry is often neglected or underestimated ... (Verkuyten "Symbols," 268-9)

Given these comments, there are at least four ways in which we might interpret and understand symbolic actions. The first is to treat them (as Nozick does) as symbolic utility vehicles (SUVs) and to then assert that an agent performs symbolic actions just in virtue of their utility. A second approach is to consider symbolic actions as analogous to written and spoken language. The third way is to treat symbolic actions as shared social practices which achieve their status cognitively through collective intentionality. (This is John Searle's approach which I consider at length in the next chapter.)

But none of these three approaches explains why people do perform symbolic actions. Simply arguing that symbolic actions have utility, and then treating symbolic utility as another form of causal utility does not explain why symbolic actions should have utility at all. The reasons why a given action (eating, for example) has a given causal utility is obvious enough as it is understood under narrow rationality, but is not so clear, on this account, why any act should have a symbolic utility. Moreover, the fact that linguistic symbols have a connotative power explains in part why language can serve a communicative function, and we can understand why we might want to communicate with others. But people do not typically perform symbolic actions for the purpose of communicating with others, so the analogy with language does not explain why people undertake symbolic actions. So the connotative function of symbolic actions is by itself insufficient to explain the existence of widespread symbolic action. And Searle's approach brings us no closer ot understanding the property of symbolic actions either. While his approach explains how social institutions such as money can acquire objective causal powers (and therefore utility) through the shared subjective will of a community that, it does not so clearly explain why people attach any utility at all to symbolic actions (though it does explain how people can share some common understanding of a given symbol). But the reasons why a symbolic action should have any utility whatsoever do not seem as clear as the reasons why other actions (seeking wealth, security, food, etc.) have utility. The fourth approach, defended here by Sperber, is to consider the emotive role of symbolic actions, and to explain why the emotive role is a sufficient explanatory foundation for symbolic utility.

Making the disanalogy between language and symbolic action clear now paves the way for Sperber to make a much stronger claim: symbols have no meaning.

"Symbols are not signs. They are not paired with their interpretations in a code structure. Their interpretations are not meanings (85)." And, Sperber says, there are enough clues for this. Symbols are tremendously uneconomical; their manifestations are vastly disproportionate to their purported meanings (3-8). Sperber claim that the "meanings" of many symbols cannot even be expressed *salva sensu* in language, and that "...it is in fact impossible to circumscribe the notion of meaning in such a way that it may apply to the relationship between symbols and their interpretation (13)."

Further, Sperber's informants are frequently unable to tell him, with any degree of certainty, just what a given symbol means, and yet their symbolic system functions very well without a complete and detailed exegesis. Nor is this state of affairs rare, claims Sperber, for we are no different in thinking that

> ...it is polite to stand up when a woman enters the room, to hold one's knife in the right hand, to cover one's mouth when yawning; impolite to point at someone, to keep one hand under the table, to pick one's nose in public. But what exactly do these different actions represent? The commentary is hesitant when one solicits it. Must we therefore say that these actions mean politeness or their opposites? Just as well to say that symbols, when they are not otherwise explained, mean 'the custom' thanks to which one avoids explaining them (21).

But interpretations can still be offered: objectors might contend that the meaning of these actions are known, but only to a few; or that the meaning lies hidden in our unconscious. It is only a matter of digging deep enough. But, as Sperber points out, the logic or motivation which supposedly informs the connection between a symbol and its interpretations in these frameworks is (as all admit)

arbitrary, just as the connections between words and their meanings are: *post hoc*, one can tell any story one wishes which explains why this should symbolize that.

The cross might symbolize Christianity because (this is the motivation for the interpretation) Christ died on it. But this answer does not explain why Christ's suffering could not be represented by the nails or his crown of thorns. Nor does it explain why the cross could equally as well represent the crimes of the criminals who also hung on it. There is no logic, no generalizable principle, which allows one to predict how people might use a symbol. In the next chapter, I explore John Searle's theory of constitutive rules whereby social facts such as symbolic action acquire their status via collective intentionality. But Searle's account, while it offers a plausible ontological explanation of social phenomena such as symbolic actions, cannot, in this case, provide their motivation. Nor can it predict what interpretation people will attach to symbolic actions.

Thus, to give a motivation for an interpretation, says Sperber, is not itself a meta-symbolic insight; the motivation is itself symbolic and must treated as such (26ff). Since the connections between symbols and their interpretations are arbitrary, any motivation is itself open to interpretation. The Usage, for example, symbolically categorize the eagle as a land animal. Why? Because eagles are associated with lightning, and lightning with fire, and fire with coal, and finally coal with the earth (26). So this motivation itself becomes symbolic of the ways in which the Usage conceive of their world, and this interpretation does not remove one from the realm of symbolism.

This, however, is not the chief defect of either the cryptological or Freudian views of symbolism. On the cryptological view, symbols originally had a

meaning, but their practitioners have forgotten it, or their ancestors forgot to pass it on. Nonetheless, a complex symbolic system of this sort can function quite well even if no exegesis is available. Even so, the determined anthropologist will seek out the experts who can provide interpretations for the symbols. But this "key" is neither needed (since many adherents may be unaware of it) nor sufficient, since the motivation for any interpretation for a given symbol must also be interpreted. On the Freudian view, the meanings of symbols may not be present consciously, but they live nonetheless in the unconsciousness of their practitioners. The real failure of both systems of interpretation, according to Sperber, is that they consider that the symbol exists prior to symbolism itself, and that they believe that to interpret the symbol is to understand both it and symbolism itself. But this, according to Sperber, is an illusion. "The notion of a symbol is not universal but cultural, present or absent, differing from culture to culture, or even within a given culture ... The attribution of sense is an essential aspect of symbolic development in our culture. Semiologism is one of the bases of our ideology (50, 83-4)." Symbolism, on the other hand, is universal. Moreover, interpreting a symbol does not explain it. For example, some Dorze rituals demand placing butter on one's head, and this might invite a Freudian explanation:

> Suppose that the ethnographer, having translated 'butter on the head' by 'semen on the genitals' takes to his heels and says, 'I have understood.' What exactly has he understood? What makes the fact of symbolically putting semen on one's genitals during certain public rituals more comprehensible than the fact of actually putting butter on one's head? The problem of interpretation is modified - as in the case of any association - but it is in no way resolved (45-6).

What then is the true nature of symbolism? Sperber thinks that humans have a universal and innately endowed, but culturally modified, symbolic mechanism which processes symbolic input. As we have seen, the mark of the symbolic act is that it is narrowly irrational - irrational, that is, in the sense that it does not seek to maximize causal utility; or, if it is a declaration ("This wine is the blood of Christ"), that it cannot be true or cannot be meant in just the way that it was said. That symbolic actions are irrational in this way may escape us when we consider that such commonalities as weddings and funerals are symbolic events. But insofar as a rational, but nonsymbolic, species would understand them, they are. Since symbolic events, by their very nature stand apart and outside the ways in which we ordinarily live, they mark themselves as special, unusual, and exceptional, and serve to focus our attention - Sperber says symbolic acts are "put in quotes (123)." Unlike language, then, there is no particular set of objects, events, or actions that count as symbols: what sets symbols apart from other objects or events, according to Sperber, is merely some marker that there is no rational interpretation for them.

Since symbols violate the canons of what I have dubbed narrow rationality in chapter II, our conceptual representations of them also fail to be subsumed under our usual modes of understanding, forcing our minds to cast about widely for any sort of connection which can reconcile them. Like a smell dimly remembered, or a snatch of music, or a particularly vibrant figure of speech, symbols are evocative - they do not lead our minds not to a single and predictable interpretation, but to a myriad of thoughts, and from those to others, to anything that might make sense of the symbol, all intended to provide some explanation of the symbolic actor's communicative intent (85ff).

Whether or not the psychological mechanisms which interpret symbols are as

Sperber describes them, I think that he is clearly right in saying that symbols derive their tremendous emotive power not from any monolithic (if unknown) meaning, but from their evocative qualities.

It is precisely because symbols are underdefined that they can evoke different but overlapping responses in so many people. And here we may begin to understand just why it is that people undertake symbolic acts: where symbols evoke a certain kind of belief or memory, they may also evoke pleasurable emotions, and these emotions - like the pleasurable emotions we feel when we perform other actions - will motivate us to participate in symbolic actions.

Evolution and the Evocative Power of Symbolic Action

I wonder men dare trust themselves with men.[27]

The next question is why symbolic actions should evoke these emotions (i.e., the ones I described in the very last paragraph) at all. Four possibilities seem to arise here.

1. A capacity for and a disposition to symbolic behavior is a non-adaptive trait (as freckles or differences in eye color might be) which neither increase nor reduce fitness.
2. They are the unavoidable side-effects of some other trait (the way in which communicative symbolism supervenes on language, perhaps?) which does enhance fitness, much as a genetically endowed susceptibility to sickle cell anemia is an unavoidable effect of increased

[27] Shakespeare, Timon of Athens. (c.1607). act 1, sc. 2.

resistance to malaria.

3. Symbolic actions are simply so useful in any set of social arrangements that all human societies must sooner or later discover their values. The employment of symbolic actions is, in short, a forced move, the existence of which requires no adaptive explanation.
4. The disposition and capacity for symbolic action are themselves adaptive and have served some important role in human evolution.

I am inclined to (4), on the grounds that (a) many forms of symbolic activity consume immense amounts of human energy, time, and other resources which, in many cases, clearly impairs survival; (b) if there was a genetic disposition to symbolic behavior and this behavior decreased fitness (as (2) suggests), then any mutation which elicited some other psychological mechanism which would override symbolic impulses would be favored (but, given the universal nature of symbolic behavior, there seems to be no such overriding mechanism), (c) many symbolic activities are clustered around central human concerns (birth, death, mating, food, and power) which are crucial to survival, and (d) it is implausible that symbolic action is a forced move as (3) suggests since performing symbolic actions only makes sense if one has a set of emotions that are triggered by symbolic actions. Other forced moves (such as using stone for tools) are forced precisely because the physical nature of our environment makes them inevitable. But symbolic actions are not an obvious response to any external features of the environment. They draw their motivation from human emotion, and it is ultimately the presence and nature of these emotions which need to be explained in order to explain the existence of symbolic action itself.

The question then becomes: exactly what adaptive role does symbolic behavior

play?[28] As I discussed at length in chapter II, Nozick considers that the primary role of symbolic behavior is its assistance to individuals in overcoming temptation, and that it may have been preserved by natural selection for this purpose.

A second suggestion, which Wes Cooper and I have recently defends, is that symbolic utility acts as a tiebreaker by allowing an agent to weight one of two equally appealing alternatives ("Buridan's Ass").

Situations of this sort are called "Buridan's Ass" Problems, in honor of Jean Buridan (1300-58), to whom the story is classically attributed: an ass, standing midway between two equally tempting piles of hay, can find no sufficient reason to prefer one over the other and , since rational action always requires a sufficient reason, the rational brute starves to death. It was also well-known to Al-Ghazali of Baghdad (1058-1111 CE). Dante Alighieri offers this view:

> Before a man bit into two
> foods equally removed and tempting, he
> would die of hunger if his choice were free;
> so would a lamb stand motionless between
> the cravings of two savage wolves, in fear
> of both; so would a dog between two deer;
> thus, I need neither blame myself nor praise myself
> when both doubts compelled me equally:
> what kept me silent was necessity.[29]

---

[28] In response to a query, I point out that this is not primarily a philosophical question. It can better be described as a question about adaptation.

[29] Dante, Paradiso, Canto IV, 1-9. Translation by Allen Mandelbaum. Cited in Skyrms, 63.

The paradox of decision under indifference is precisely that such decisions seem rationally insoluble, and yet in practice they typically offer us no problem at all: we do not find ourselves volitionally frozen when presented with qualitatively identical books, apples, or what have you: we simply take one or the other without much regard for choice. But how is this possible and how can it be defended within a theory of rationality such as Nozick's? One suggestion is that we have a randomizing mechanism that chooses one option for us, as it were. But this option, says Brian Skyrms (64), only opens up a regress of nested nil-preference problems: how do we choose between two equally appealing randomizing mechanisms? Even if evolution preempted our choice of a randomizer by simply endowing us with one particular randomizer, there would still remain the problem as to how the human brain could interpret the output of a general-purpose randomizer into a decision to go left or right, up or down, or whatever. And we would also be at a loss to understand why the output of a randomizer would count as a reason to act. Given that we want to solve the problem, and that we are not adverse to using a random process to make the decision for us, we can still wonder why we should be motivated to adhere to the randomizer's output. Why not its contrary? And so on.[30] Cooper and I therefore argued that humans resolve Buridan's Ass problems by simply telling ourselves, as it were, that we prefer one alternative over the other "just because we feel like it" and that this reason is sufficient reason for us to act because of the importance that expressing ourselves symbolically in this "just because we feel like it" way plays in our emotional lives.

[30] Satisficing doesn't seem to be a meaningful alternative to maximization here, because the conditions of the Buridan's Ass problem imply that both choices will maximize utility and that both will satisfice. Satisfying simply asks the agent to take the first adequate option available. But in Buridan's Ass problems, neither option is prior to the other in any way. The Buridan's Ass problem is a problem in indifference, not maximization.

Whether or not these are adequate accounts of overcoming temptation or making decision under indifference is not the central question here. It seems more pertinent to our purposes to ask whether either line of argument will count as a sufficient explanation of the adaptive role of symbolic utility. I have to answer in the negative for two reasons. First, overcoming temptation and making decisions under indifference by appeal to symbolic utility are both cases of an agent imposing symbolic utility on acts which would otherwise be rationally acceptable options. (i.e., choosing them would not decrease an agent's utility, even though it is true that, before imposing symbolic utility on them, the agent finds no sufficient reason to perform them.) But many symbolic actions are not of this sort. In these cases, an agent does not simply assign a symbolic utility to some act that she might have performed for some nonsymbolic reason. Rather, choosing a symbolic act requires her to enter a symbolic realm, where the option of undertaking an act exists only in virtue of its symbolic utility. Symbolic acts of this sort are typically elaborate and ritualized. They are not choices which which exist antecedently to our symbolic motives, they are instead created by our symbolic motives. And I think these sorts of symbolic actions form the great majority of symbolic acts.

Second, overcoming temptation and deciding under indifference are problems in parametric choice, but symbolic acts reach their highest form of expression[31] as social acts. Our richest and most diverse symbolic acts derive their importance from their roles in marking social, rather than purely personal, events, values, and commitments. Accordingly, it seems more probable that the origins of human symbolic behavior lie in its social role, and since humans are the only species who engage in such elaborate symbolic behavior, the explanations for

[31] By "highest form of expression" I mean most elaborate, ritualized, tradition-bound, intricate, costly, and most fraught with emotion.

this must lie in the specific trajectory of human evolution. Specifically, the highly social nature of the lives of ancestral humans has exerted a strong selective force on our psychology and set a premium on behaviors that enhance group cohesion and that allow individuals to "self-announce" their loyalty to the group and to other individuals.[32] Symbolic action, I believe, plays just such a role. The next section lays out a line of anthropological and psychological evidence to support this contention.

The Evolution of Intelligence

The evidence for when and how various aspects of modern human behavior - intelligence, language, symbolism, and culture - evolved is still a deeply controversial - and perhaps intractable - problem in contemporary anthropology. For one thing, any such account must necessarily recognize that the changes in early human cognition took place within a complexly interrelated nexus of causal factors - environmental changes, global dispersal of *Homo*, dietary changes, changes in seasonal migration, changes in hominid group size and hierarchality, decreased sexual dimorphism (chiefly in body size), increased manual dexterity, the advent of bipedalism, prolonged maturation, increased encephalization, increased neocortex size, the development of fully modern vocal systems, and so on - some of which drove cognitive change, and others which were dependent on

[32] Does this mean that Robinson Crusoe or other solitary humans wouldn't use symbols? The fact that a given adaptation evolved in one context does not in any way mean it cannot or would not be used in another context. An example of a non-psychological adaptation will make this clear. The bodies of many monkey species are adapted for an arboreal life, but this does not preclude them from exploiting their physical traits in non-arboreal settings (e.g., cages).

it.[33]

Secondly, the chief analytical tools by which anthropologists attempt to understand the evolution of the human mind - inferences about technological capacity from archeological finds, inferences about likely intellectual capacity made from the cranial capacities of hominid fossils, analogies from surviving hunter-gatherer societies, analogies from the behavior and biology of other surviving primate species, and inferences from the development and growth of individual human beings ("ontogeny recapitulates phylogeny") - do not give any direct indications in themselves of either the mental capacities or the actual behavior of early humans.[34] Rather, each analytical method should be used circumspectly and in conjunction with the findings from other fields. As an outsider, it appears to me as if there is still considerable room for informed speculation. That said, let us see what the past can tell us.

Homo habilis, the first distinct member of the genus, first appeared about 2.4-2.0 million years ago and "modern" *Homo sapiens* appeared about 130-100,000 years ago. On the basis of archeological evidence, these early humans made very slow cultural and technological changes until the beginning of the Upper Paleolithic, about 40 or 50,000 years ago. "Prior to this time, human morphology and behavior evolved slowly, hand-in-hand. Afterward, fundamental morphological evolution all but ceased, while behavioral (cultural) evolution accelerated (Klein, 190)." This turning point in human history has been variously named the

[33] Why should our understanding of the development of human behavior attempt to recognize all these factors? Because many changes have multiple causes, and scientists have found that explanations which take into account all the relevant factors (or as many as is possible) are more likely to be closer to the truth than those that ignore relevant factors.

[34] With the exception of archaeological finds of tools, ornaments, art, habitation, etc. that clearly do display some aspects of human behavior.

"linguistic," "cultural," "software," or "symbolic" revolution - and is marked by a large number of profound changes in human behavior (Klein, 168; Mellers, 63). These changes included: increasingly standardized, regionally diversified, specialized, and economically produced stone tools, increased tempo of technological change, first creation of relatively complex bone and ivory artifacts, personal ornaments, art, ritual, and increasingly sophisticated methods of social organization (Mellars, 63-5). By this time, early humans had dispersed across Africa, Europe, and Asia, and were living as hunter gatherers in widely differing environments and in groups that were more adaptable, mobile, egalitarian, and likely larger, than the social groupings of their hominid forebears. Although the Upper Paleolithic provides the earliest unambiguous evidence of symbolic behavior, it does not provide evidence that there was a change in human capacity at that time (Renfrew, *passim*). For example, the archeological records left by computer users, modern hunter gatherers, and humans who lived 10,000 years ago would not reveal that each group had roughly equal cognitive capacity. Given that there is no obvious morphological change between Middle and Upper Paleolithic humans (excluding Neanderthals in Europe), humans may have had these capacities since the emergence of modern *Homo sapiens* some tens of thousands of years before, even though the archeological record does not provide direct evidence of the existence of these capacities.

So our account here is speculative, as it surely must be. However, there is broad theoretical support for the notion that many aspects of rationality, especially those concerning social interaction, owe their nature to their role in promoting individual fitness in the ancestral environment and I feel that such suggestions, while not conclusive, are a useful heuristic for investigating symbolic utility and action in a way which coheres with evolutionary psychology and with evolutionary explanations as a whole.

Intelligence is notoriously difficult to define, and some commentators have argued that it is meaningless to even try to make interspecies comparisons of intelligence. Nevertheless, let us define intelligence, roughly, as the ability to respond flexibly, quickly, and appropriately to novel and changing conditions within one's environment (comprising the traits Jack Copeland subsumes under "massive adaptability" (Artificial Intelligence, 55)). Defining intelligence in this way makes its adaptive powers more obvious (albeit vaguer) than if we define it by, say, the ability to achieve a certain score on an I. Q. test. But some of the discussion to follow will help to clarify our understanding of human intelligence. Moreover, intelligence thus understood relies heavily on an individual's ability to abstract relevant information from a multitude of sources, make reliable inductive and deductive conclusions, to - in general - be responsive to reasons. So our definition of intelligence may stand as a rough measure of an individual's capacity for rationality.

Two questions arise here: (1) why did humans - and to a lesser degree other primates - develop intelligence at all, when so many other species have managed to flourish with a limited number of more or less fixed responses to environmental changes? (2) why do humans - and again to a lesser degree, other primates, especially chimpanzees - possess intelligence which is capable of performing tasks that are both vastly different in nature and more complex than those they would have encountered in the EEA? In the ten thousand years or so since humans have largely abandoned hunting and gathering, there has been little, if any, increase in our cognitive powers, but our technological capacity has increased many thousand fold.

These are not idle questions, and I offer some answer to them shortly. Primate,

and especially human intellect, is costly. While there is no methodologically uncontroversial way to correlate intelligence to absolute or relative brain size between primates, it is obvious that humans have developed intellect by increased brain size as well as by brain reorganization. But large brain size entails several costs: long pregnancies while the fetal brain grows, relatively painful and risky childbirth's (since the average infant head is slightly larger than the average birth canal), long postpartum development while the brain develops even more, and high energy consumption (the brain, while comprising only 2% of the body's weight, consumes 20% of its energy budget) (Berkow et al, 279).

Since, as Nicholas Humphrey ("Intellect") notes, nature ruthlessly prunes traits which are costly to the organism while not enhancing its fitness, there must be some biological function or set of functions which intellect performs which justifies these costs. There are some hints as to what that might be. Alison Jolly notes that

> ... learning is not a generalized ability; animals are able to learn some things with great ease and others only with the greatest difficulty. Learning is ... the process of acquiring skills and attitudes that are of evolutionary significance to a species when living in the environment to which it is adapted. (Washburn et al, cited in Jolly "Lemur Social Behavior", 28).[35]

For example, a monkey that may require lengthy pre-training and adaptation to an apparatus as well as 20 to 100 trials to solve one two-choice object-

[35] It might be objected that I have already defined intelligence as the ability to adapt to new environments. But there is no contradiction between Washburn and me, since neither of us are saying that either adaptation or the learning processes that it favors will create intelligence that is truly domain-general. So humans may display intelligence in adapting to some new environments, but they may also fare poorly in others that are more dissimilar to the EEA.

discrimination problem will, in a matter of seconds, or, at most minutes, become thoroughly introduced for the first time to a social situation with three or four cage-mates. (Zimmerman and Tory 1965, cited in Jolly, 1966/1988, 28).

Jolly observes that lemurs direct learning and insight towards three areas of adaptive interest: objects (including food), other animals (chiefly predators), and conspecifics.[36] But, she argues, it is this latter area which is of fundamental importance and therefore the best measure of lemur intelligence, because by learning sociability from conspecifics, a lemur best equips itself with the capacity to deal with all three areas. Therefore social integration and intelligence are mutually reinforcing and would have co-evolved (29-30). Because lemurs, like other primates, can learn from each other, they need not have well-developed capacities specific to investigating or learning about objects.

Humphrey suggests that the challenges of primate life are far more social than technological. While Humphrey does not deny the importance of rudimentary technology to chimpanzees and humans, he contends that the requisite technology can be achieved by "low-level" intelligence (16), relying more on trial and error, serendipity, and imitation, than on serious and prolonged inductive investigation. This effectively reduces the individual's need for what Humphrey calls "practical invention", since the individual, instead of solving every new environmental problem for herself, can benefit by imitating the most effective solution to parallel experiments performed by other individuals.

But this approach presupposes a close degree of social cohesion which imposes its own costs. Specifically, the demands of social interaction - of being able to

[36] Jolly is simply noting that the aspects of a lemur's environment that are relevant to the lemur's psychology can be fitted into these three categories. I do not read her as making all-embracing claims about the totality of nature.

maintain an accurate mental representation of the dispositions of each member of the group, and to anticipate how one's actions are likely to affect others - is far more mentally demanding. Social problems are dynamic in ways in which nonsocial problems are not - because the other elements in the situation have their own interests which may or may not coincide with an individual agent's. That is, while individuals will be concerned to acquire food, mates, personal security, and to protect near kin, even if this sometimes deprives other group members of scarce resources, they will also be concerned to protect the social nexus, since this is what makes their existence possible.[37] "Someone embarking on such a transaction must therefore be prepared for the problem itself to alter as a consequence of his attempt to solve it (Humphrey, 23)." And in such situations, there will be strong selection for a large suite of social skills. In short, wherever social success affects fitness, and where the capacity for social success is heritable, then organisms with greater social success will increase their fitness and this adaptation will cascade through an entire population (although differentially).

Humphrey's chief conclusion is that the propensity for social thinking is so strong that humans have approached a wide number of recalcitrant natural phenomena just as if they were social phenomena that can be solved by social transaction - religion, witchcraft, and animism being three very obvious examples of our efforts to argue with nature, rather than to think <u>about</u> nature.

---

37 An expansion of this argument can show why one central dispute between liberals and communitarians is particularly sterile and unhelpful: communitarians claim that society is ontologically prior to the individual, and that individuals are created by society . They further contend that liberals, who claim that the individual is ontologically prior to the group, must believe that individuals must therefore form all their desires outside the group (whatever that might mean). Humphrey's analysis shows that groups are only possible if individuals have some preformed capacity for social interaction (that is, even if social groupings do form individuals, they cannot do so with just <u>any</u> individual - it must be a biological individual who already has a capacity for social interaction (and this capacity implies a broad set of social-cognitive skills), and human individuals are (generally) only viable if they are in groups.

The rise of scientific method, he thinks, is an effort to overcome this propensity (22-6).

Jolly's and Humphrey's work has consolidated into what has been dubbed the Machiavellian intelligence hypothesis (Byrne and Whiten, 1-10). This theory is founded on the premise that in many species, individuals have interests which sometime conflict and sometimes coincide. They are thus in incomplete social harmony. In some cases, individuals can cooperate with each other without fear of decreasing individual fitness. But some kinds of social cooperation require altruism - that is, one individual can only confer a benefit on another by (temporarily) suffering a loss to her own fitness.[38] For example, warning other conspecifics of a predator may endanger the individual or food-sharing may not be reciprocated. Nonetheless, groups that practice reciprocal altruism may in many cases fare better than groups who do not. But this fact will not be sufficient to ensure that an altruistic strategy will prevail over defection, since defectors within a group of altruistic cooperators will fare better than other group members and may quickly subvert altruism from within, leading to widespread defection, and this prediction is mirrored exactly in the conditions of the Prisoner's Dilemma. To be selected, therefore, altruism must not only contribute to group fitness, it must also contribute to the fitness of the individual (Sober Biology, 86).

However, the conditions under which reciprocal altruism can arise are rare. Douglas firs, for example, expend vast amounts of energy competing for sunlight by achieving great heights because the have no way of "agreeing" to mutually

[38] It is important to note that altruism, described in this way, carries no connotation of moral behavior and does not rely on any assumption that members of other species have any moral dispositions.

restrict their height. Organisms that engage in reciprocal altruism require a set of traits need to make such exchanges mutually beneficial while minimizing (or reducing) the threat of free-riders. G. S. Wilkinson shaped his successful research into food sharing between female vampire bats by predicting that female vampire bats, on the other hand, engage in mutual blood sharing because they have a specific set of psychological traits :

> I needed to demonstrate that five criteria were being met: that females associate for long periods, so each one has a large but unpredictable number of opportunities to engage in blood sharing; that the likelihood of an individual regurgitating to a roostmate can be predicted on the basis of their past association; that the roles of donors and recipient frequently reverse; that the short term benefits to the recipient are greater than the costs to the donor; and that the donors are able to recognize and expel cheaters from the system. (77; cited in Barkow, Tooby, and Cosmides, 169)

Since humans (and other primates) have engaged in social exchange for long periods of time, it follows that we too must have acquired a set of traits specific to the task of enforcing cooperation and punishing cheaters (Tooby and Cosmides offer an extensive list of the necessary design features necessary for successful and continued social exchange, 177ff). For example, John Tooby and Leda Cosmides argue that humans do not have a general-purpose ability to detect violations of conditional statements ("if P, then Q") and in fact frequently cannot consistently solve such problems. On the other hand, Tooby and Cosmides demonstrated that individuals could accurately detect violations of conditional social contract laws ("If a person is drinking beer, then he must be over 20 years of age"), and by eliminating other possible explanations, they were able to show that the reason for this is due to the existence of a domain-specific,

species universal, innate cheater detection mechanism (See Cosmides and Tooby "Cognitive Adaptations" *passim*, for extensive discussion of the experimental framework and the elimination of competing hypotheses).[39]

Machiavellian intelligence also implies that individuals will benefit if they can engage in an "arms race" of increasingly elaborate strategies of defection, deception, detection, punishment, and so on. For example, Jolly ("Communication," *passim*) and Andrew Fenton review research showing that chimpanzees frequently engage in tactical deception by hiding food from others, leading others to places where the hidden food is not located, hiding undesirable food in a way that distracts others while they retrieve other, more desirable, food, etc. In one case, a young male concealed his erection from a dominant male (who would have punished him for the display) while revealing it to females whom he hoped to entice. In another instance, a pair of male chimpanzees deliberately hid facial expressions that they could not control so that the other chimpanzee could not see them.

The Machiavellian hypothesis, however, does not entail that all members of a social group will be evolve to be cunning and self-interested opportunists who only pretend to cooperate but who in fact care nothing for the others and are greedily awaiting the slightest chance to get an adaptive leg up on them.

---

[39] It has been suggested that the "cheater detection" is somehow equivalent to, or can be translated as,"conditional processing." But Cosmides' and Tooby's work does not apparently support this contention, since the pair think they proved that the two are not equivalent. As they summarized their own findings: "Virtually all the experiments reviewed above asked subjects to detect violations of a conditional rule. Sometimes these violations corresponded to detecting cheaters of social contracts, other times they do not. The results showed that we do not have a general-purpose ability to detect violations of conditional rules. But human reasoning is well designed for detecting violations of conditional rules when these can be interpreted as cheating on a social contract ("Cognitive Adaptations", 205)." Since humans have a cheater detection mechanism, but not a generalized ability to to detect violations of conditionals, it would seem that the two cannot be equivalent.

Deception is costly and difficult to maintain. At least in the case of humans, it seems at least possible that individuals may fare better when they are sincere cooperators, and have strong internal motivations to cooperate consistently. Thus, the Machiavellian hypothesis also explains the existence of emotions - liking, anger, gratitude, sympathy, guilt, shame, and so on - in terms of their functional role in creating and enforcing stable patterns of social exchange that benefit their possessors (Pinker, Mind 403-5).

The Adaptive Value of Symbolic Action

I propose that a large and central class of symbolic actions are performed because they are evidence of an agent's deontological commitment to some set of actions that will typically be directed towards adaptive sub-goals (acquiring food, mating, reproduction, protection of self and offspring, building alliances with others, resisting aggression from others, etc.) - even though the agent need not - and typically is not - motivated to perform these actions specifically for this reason. These symbolic actions include initiations and rite of passage rituals, expressions of friendship, gratitude, contempt, contrition, fealty, and romantic love, displays of personal courage, revenge, some forms of punishment, etc. Following Humphrey, I further propose that self-directed symbolic actions and (such as New Year's resolutions to stop smoking, etc.) and symbolic actions directed towards nature or supernatural forces (including gods and spirits of the dead) can be understood as extensions of this first, social, application of symbolic

action.[40]

Symbolic actions can perform this self-expressive and self-announcing function because they are, by their very nature, recognized as such because they are narrowly irrational. That is, agents who engage in symbolic action impose a cost on themselves and it is this cost that in part renders their acts symbolic. In so doing, they signal to the community at large that they are not simply maximizing causal utility, but are acting in defiance of it. An individual who was simply pretending to announce his commitment to others, as a way to secure cooperation from others, would be conspicuous in his calculation of just how to signal his commitment in a way that maximizes utility.

Brian Skyrms has modeled a problem in game theory that vividly illustrates both the magnitude of the benefits of even moderate amounts of self-announcement and the degree to which self-announcement favors only certain interactive strategies. Imagine a game in which pairs of agents independently choose what fraction of a cake they are individually willing to accept. They then receive that share iff the sum of their claims sum under unity. An indefinite number of solutions is possible, but if both agents wish to increase (but not necessarily to maximize) their share of the cake, they may not be able to solve the problem. This simple game brings out the tension inherent in many social interactions in which individuals are in imperfect harmony with each other, having reason both to cooperate (since too much greediness makes them both lose) and to exploit the other (since one can achieve a gain only by inflicting a loss to the other). Now

[40] It might be objected that when people in engage in symbolic actions directed at supernatural forces, they typically do so because they think that these actions have the direct and desirable causal effect of influencing the gods in some way. And therefore people are not performing these actions only because they have some symbolic utility that redounds to their own personal well-being. But this begs the question: given that one wants to to thank, appease, or praise the gods, why would one think that symbolic actions are an appropriate way to perform these actions?

consider a population of such agents who pursue three different strategies in a series of such cake-sharing games played against randomly selected opponents. A Greedy demands 2/3 of the cake and is doomed to lose whenever she encounters another Greedy. Fairminded demands 1/2 and therefore fares well with her own kind, but gets nothing when confronting a Greedy. Modest asks for only 30% and therefore always receives her share of the cake no matter who she plays against. - but she always fares worse than Fairminded and Greedy when she plays against them or other Modests (12-3). Successive generations of agents inherit their strategies from their "parents" (the agents who play against each other pairwise) and the percentage of offspring from each parental pair who practice a certain strategy is equal to the relative proportion of the cake that that strategy obtained for the parent in the previous generation (11).

Given these constraints, populations composed entirely of Modests or Fairmindeds are stable, but a population composed exclusively of Greedies could not survive. However, initial mixes of agents that are heterogeneous (containing varying proportions of Modests, Fairmindeds, and Greedies) will settle into one one of two stable equilibria: an "egalitarian" equilibrium composed entirely of Fairmindeds or an "exploitative" equilibrium state composed of equal parts Greedy and Modest.

We can now complicate this picture by imagining that the players do not choose their opponents at random, but are instead more likely to play against like individuals. Greedies will lose from this strategy because they will more frequently encounter their own kind, Modests will fair no better or worse than before, but Fairmindeds will benefit overall by avoiding some destructive encounters with Greedies. Skyrms ran computer simulations in which Greedies,

Fairmindeds, and Modests played series of games in just this way. He found that if there is even a small degree of correlation between types of players (say, 0.10), the Fairminded strategy is far more likely to go to fixity, and if the degree of correlation is 0.20, any initial population mix of all three strategies will evolve to 100% Fairminded (15-20). This is because only Fairmindeds stand to benefit from recognizing their own kind.

The real world analogue of this model is complicated by the obvious problem that Fairmindeds cannot benefit from correlation unless they can recognize each other as similar, and that (if we vary the game somewhat) Greedies can profit from presenting themselves as Fairmindeds.[41] For example, as we saw in chapter II, players in the Prisoners' Dilemma are rational to cooperate in a one-shot PD if they can recognize each other as similar. The defector (or Greedy) who pretends to be a cooperator (Fairminded) can thereby lull her cooperative opponent into an unfavorable outcome. Given plausible assumptions about our ancestral life - that social exchange was common, that individuals were not uniformly cooperative, that some exchanges entailed that one party had to trust the other, that some of the exchanges involved nontrivial utilities, that parties could therefore profit or lose substantially from deception, and so on - there would have been strong selective pressures for mechanisms that would have allowed cooperators/Fairmindeds to identify each other with some degree of reliability. And, given Skyrms' results, this mechanism need not have been infallible, since only a modest degree of correlation via self-announcing yields a high adaptive advantage. I am suggesting here that symbolic action provides just such a mechanism, and this chapter will explain why.

[41] This isn't of course possible in Skyrms' example, but I am now describing an application of his model to the real world.

But how specifically and exactly could self-announcement through symbolic action confer an adaptive benefit on an individual? Here is an example that I think nicely illustrates this. David Buss argues that humans who were indifferent about mate selection would have less fitness than humans who were carefully selective. On the other hand, all other things being equal, humans who seek mates whose physical and behavioral characteristics and reputation correlate highly with their likely future success as mates will increase their fitness ("Mate Preference Mechanisms" 92-3). But to accomplish this, an individual need not consciously calculate who is the best available mate. It will be sufficient if individuals merely have some cognitive mechanism that can determine the best available mate and can then trigger some affective state (call it "falling in love") that will motivate the individual to pursue the object of his/her affections. This account does not deny that emotions play an important causal role in sexual attraction and mate selection, but it insists that the emotions considered by themselves do not explain human behavior, since it is frequently the emotions themselves that need to be explained.[42]

The problem is that what counts as the best available mate now need not always be the best available mate forever. Over time, an individual may come to believe that some other individual is more desirable than her current mate. And precisely the same rational, fitness enhancing considerations that caused her to pick her first mate may now motivate her to abandon that mate in favor of another. But precisely the same considerations apply to her current mate, and neither is rational to mate with a mate who is likely to subsequently abandon him/her. The only way out of this paradox is for each partner to convince the

[42] One objection to this account is that is is a "just-so" story. This objection would have considerable force if I was defending this account as an explanation for the existence of romantic love and behavior associated with it. But this is merely an example intended to illustrate that apparently irrational action of a certain sort could enhance fitness and therefore be selected for.

other that his/her commitment to the other is not subject to reconsideration no matter what future temptations arise. So this "arms race" demands that individuals convince prospective mates that they are not in fact self-interested defectors, and this, argues Stephen Pinker, is why protestations of love are so frequently about how the wooer is crazy, smitten, bewitched, et cetera (Mind, 417-9). Another way to convince someone else that one is firmly committed to deontological duties of one sort or another, with no chance of being swayed by future utility-maximizing considerations, is to engage in symbolic activity (buying elaborate gifts, etc.) since such activity, by its very nature, announces itself as irrational.

Is this a plausible claim? It is insofar as it conforms with the predictions of game theory and insofar as game theory has proven to model adaptive behavior in the real world. It will derive additional support to the degree that evolutionary psychologists can show that there is a strong correlation between behavior that this model posits and predicts as fitness-enhancing and its actual contribution to fitness in human populations. Evolutionary psychologists can also support this claim by showing that human sexual dispositions are sensitive to environmental variations. For example, some evolutionary psychologists argue that some species-typical indicators of human beauty (clear skin, symmetrical features, evidence of youthfulness, etc.) are rough indicators of physical health and this explain why humans prefer attractive mates (as measured by these criteria) to less attractive ones. But physical health is only consideration when picking a mate and it should be less important in areas where disease is rare than in areas where it is prevalent.

Humans do not benefit only when prospective mates recognize them as reliably faithful. In any human society, individuals will secure significant benefits when

others trust them and are willing to help them on the assumption that this help will be reciprocated and here too symbolic actions will serve to announce their social reliability. Conversely, an agent will be exploited if others recognize her as one who always cooperates and who does not resent it when help is not reciprocated. But they will be less likely to override the interests of an individual who is willing to avenge any challenges to his/her well-being even when the vengeance may cost more to the agent than the immediate loss that motivated it. Margo Wilson and Martin Daly argue that a large class of homicides in the United States are inexplicable unless we interpret them in this way. These homicides typically involve males who kill each other over apparently trivial provocations - what Wilson and Daly call "slight or offense." Although it appears irrational to kill another member of one's society for something as insignificant as jostling or a spilled drink (and to thereby risk severe retaliation), these actions make more sense if we see them as symbolic actions that clearly announce to others that an agent will not tolerate any violations of interests, no matter what the cost to himself. Why should he do this? Because if he allows others to get away with a relatively minor affront, they will be emboldened to escalate their attacks on him until they become very real threats to his fitness. Thus a preemptive symbolic display to deter such attacks is warranted (Daly and Wilson Homicide, 123-36).

So a disposition to announce one's willingness both to cooperate and to punish defectors will be favored by natural selection. Moreover, the actor's willingness to "pay" the costs of symbolic self-expression should correlate reliably with his dispositions to act in certain ways in the future, and other members should be able to recognize this (The two examples above are meant to illustrate - not prove - how this correlation might obtain and how other might recognize it). This fact reduces calculation costs for other individuals when they need to appraise the

actor's likelihood of cooperating or defecting. The assumption that meaningful and continued social interaction relies on agents being able to understand and predict other agents' preferences and likely actions is implicit in some interpretations of game theory in which agents are assumed to be preferentially "transparent." The fact that people are motivated to engage in symbolic actions because their emotional value (i.e., I attend a funeral, not to "be seen" but to fulfill some internal, but vaguely understood, feeling of obligation that it is pleasurable to fulfill) and the fact that spectators also value them for their emotional value is what enables them to perform this function. If both actors and spectators recognized a symbolic action as no more than the price of being recognized as a reliable cooperator, it could not, for that very reason, fulfill that function. The crucial point to remember here is that the reason (on this account) why people perform symbolic actions is because the propensity to do so served as a reliable indicator of their willingness to cooperate with others in the past, and being recognized as a cooperator contributed to fitness and a propensity to symbolic action was preserved by natural selection for that reason. But this is not people's motivation for performing sincerely motivated symbolic actions. Rather, their motivation for acting symbolically is the emotions which are associated with symbolic action. These too were favored by natural selection, but in a way that

resists, to some degree, deception.[43]

[43] Smiling might work in the same way. On an evolutionary account, the reason people smile is that, as a facial expression, it is a reliable indicator of one's state of mind - in this case, pleasure. the smile expression therefore likely served a fitness-enhancing function in the past, since people who signaled sincere happiness on appropriate occasions would fare better than those who didn't. And this is why facial expressions are species-typical and found in all societies, and why infants, even blind ones, display facial expressions from a very early age (Pinker Mind, 365-6). But this is not the motivation why people smile. People smile as a more or less involuntary expression of happiness. In these cases, they are motivated by their emotions, not by a direct interest to benefit themselves or to increase their fitness. It is, of course, possible to smile deceitfully, but this is difficult, since it involves a different set of muscles which are controlled by another part of the brain, and many people are adept at detecting fake (airline stewardess) smiles (Pinker Mind, 415). I offer this example as an illustration of the difference between a fitness-enhancing reason for the existence of a specific behavior and and the more proximate emotional motivation for its actual occurrence. Nothing in my discussion of symbolic action turns on whether this explanation of smiling is actually the correct one.

A Caveat About Sex Differences

At this point, I need to offer a caveat indicating that this reconstruction may need to be importantly modified to accommodate sex differences. My use of "men" and "women" as meaningful categories under which to group human individuals is not subject to the objection that there is no non-question-begging way to categorize some individuals. "Woman" and "man" may be fuzzy concepts, but that does not mean that there can be no meaningful statistically measurable differences between most of the members of each group.

The fact that men and women differ biologically dictates that different strategies in the EEA would have increased male and female fitness.[44] Evolutionary psychologists argue that since women invest far more in a pregnancy than men do in sperm production, the prospects for reproductive success therefore vary greatly between the sexes. A particularly successful man could father perhaps a hundred children over his lifetime, but could also (if he were not so apt) be eliminated from the reproductive game altogether. Women, on the other hand, could probably only give birth to at most three or four viable offspring, and could have best maximized their reproductive success by careful mate selection, investing heavily in their successful offspring and by inducing their offspring's' father to also invest in them. Since men's potential reproductive loses and gains are subject to greater variation dependent on ability and strategy, men would have profited from riskier strategies, greater competition, and even by facing

[44] Of course, everybody differs biologically. But this truism in no way counts against my claim that the specific biological differences between sexes (and especially those related to reproduction) may be responsible for psychological changes.

greater risks of conflict with other males.[45] Correspondingly, men may have greater concern to strike a social contract that enforced equality between them as a way to deflect the trajectory from competition to outright violence. In this way, the desirable equilibrium converging on Skyrms' Fairmindeds would have been realized as the well-documented egalitarianism that prevails in most hunter-gatherer societies (Diamond Guns, 267-70). But this egalitarianism need not have perfused relations between the sexes. If women had less reason to compete with others for resources, status, or mates as a way to increase fitness, they may also have stood to gain less by devoting resources seeking and enforcing political equality.

This is a rather abbreviated exposition of what is a complex and contentious set of claims, and I shan't review all the literature on this divisive question. Nor do I intend to argue for these or any specific adaptive psychological differences

[45] It might be objected that it is unclear why there should be any variation in reproductive success between men and women at all, and why this variation should rely in any way on ability or strategy. Women can only conceive and bear children at best once every nine months, and probably even less frequently than that in most ancestral environments. Moreover, the conditions of ancestral life suggests that women typically bore only three or four successful offspring (Daly and Wilson Homicide, 39-41). Men, on the other hand, can (in theory, at least) successfully impregnate a woman every few hours, and could father hundreds of children in a lifetime. For example, Napoleon Chagnon described one successful Yanomamo warrior named Shinbone who fathered forty three children by eleven wives, 120 grandchildren, and at least 480 great grandchildren (Daly and Wilson Homicide, 133). Since the interval between zero and one hundred is greater than the interval between zero and four, this effectively proves that potential variation between male reproductive success is subject to greater variation that the reproductive success of women. The second part of the proof must show that at least part of this variation can be explained by difference in ability. Obviously a man who is cognitively inadequate even to the task of impregnation will have less success than men who are capable of doing so. Further, men must either attract mates, buy mates, or force them to mate. In any of these eventualities, they may frequently have to compete with other males, and if they are less competent, they will mate less frequently or not at all. (See Daly and Wilson (Homicide, 136) for sources which argue that well-respected men attract more mates.) Similar comments apply to men's choice of strategy. Further yet, men may differ in their competence at providing for offspring and this will also affect reproductive success. To be sure, female competence and strategy selection also will affect differences in their own reproductive success, and where female parental investment is greater than male parental investment, this may offset some of the differences I've outlined above, but to a lesser degree, since less variation in reproductive success is possible.

between men and women. I do want to make the point that it is irresponsible to assume that there are no, and could be no, selective pressures that could have created psychological differences between the sexes, in the first place because it seems to me that there is simply no proof for such a claim, and in the second, because studies such as Daly's and Martin's present observations (following) that make this contention improbable. And I follow Daly and Wilson in suggesting that the onus for proof does not lie on those who assert the exists of any innate any innate differences between the sexes:

> If the conventional wisdom is true - that is, if the psyches of women and men have *not* been differentially shaped by selection - then someone has to explain *why* not. Why should patterns of differential reproduction have been without selective consequences in this one sphere, namely behavioral sex differences, when they have so clearly been effective in shaping those aspects of behavioral control systems that are not sexually differentiated? (Wilson and Daly Homicide, 160; emphasis in original)

And, we might add, if there are selective consequences that have been responsible for sex differences over a range of non-psychological functions, why shouldn't there be sex differences in mental function as well? Because it's *a priori* impossible? Because the fact of substance dualism makes it physically impossible? Because it's politically impossible?

The only way to settle these questions is by undertaking careful cross-cultural studies over a wide range of behaviors and evaluating the differences carefully. This is in fact is exactly what Daly and Wilson (Homicide), among others, have done. They have analyzed 35 sets of data on homicides from 21 cultural groups

in five continents ranging from the !Kung San to modern industrial societies and dating from the fourteenth century to the present. In all but one case, they found that men committed between 91% and 100% of all the same-sex homicides. The lone exception was Denmark (1933-1961), where women accounted for 15% of same sex killings, all by mothers against dependent children (147-8). These data demonstrate, they say, that "[i]ntrasexual competition is far more violent among men than among women in every human society for which information exists (161)." Wilson and Daly argue that the most plausible explanation for this difference is that because men can increase their fitness more through violence than can women, such behavior has been favored by natural selection.

Clearly, men and women may differ on some behaviors and not on others, and there is likely to be a wide quantitative overlap between the two groups for many traits. There is, most importantly, a need to avoid dogmatism on this matter. Feminists have argued cogently both that unitary accounts of rationality frequently assume (explicitly or otherwise) that "male" rationality is representative of all human rationality and that dualist accounts of rationality typically devalue "female" rationality. At this moment in time, it seems to me that confident assertions of "no differences" or "many ineradicable differences" should both be viewed with suspicion.

For the narrow area that I propose to explore, it is difficult to see any compelling reason to suspect that there is a significant difference between the way that men and women use symbols. Nonetheless, this work is written with the caveat that future research could discover differences in this area.

This is, I stress, a speculative account of how we might have come to be a symbolic species. But it is, in virtue of its consilience with the social explanation

of human intelligence, a plausible one. The next chapter will explain, in a more philosophical manner, the social recognition of symbols can be constructed from individual intentionality.

Chapter V

The Social Ontology of Symbolic Action

There is something in this more than natural, if philosophy could but find it out.[46]

The last chapter attempted to identify the most likely (or at least one possible) source and purpose of the mechanism that allows individuals to attach value to symbolic actions and that therefore motivates symbolic action and individuals' adoption of shared symbolic values (the process Robert Nozick dubs the "downward" component of symbolic utility (32) and which Paul Viminitz calls "symbolic uptake (p.c., 1999)"). This chapter attempts to explain how individual understandings about symbolic value can create social facts and social functions by subsuming symbolic value within John Searle's discussion of the creation of social facts. That is, I shall explain how entities such as symbolic action can find a place within a naturalist ontology when explained as social facts with a certain type of status function. But the claims of the last two chapters pose several challenges here.

For one thing, the naturalism that informed the evolutionary arguments of chapter III demands that any account of social entities must be solidly rooted in natural facts. "Social facts", whatever they are, must owe their existence to, and supervene on, natural phenomena. By this I mean that it is therefore not open to me argue that "Culture is a thing *sui generis* that can be explained only in terms of itself .... *Omnis cultura ex cultura* (Lowie, 66)" or that "the determining cause of a social fact should be sought among the social facts preceding it and not among the states of individual consciousness (Durkheim 110)." Rather, following

[46] Shakespeare, Hamlet, act 2. sc. 2.

Popper, I shall assume that "[a]ll social phenomena ... should always be understood as resulting from the [mental] states and actions of individuals, and we should never be satisfied by an explanation in terms of so-called 'collectives' (Popper 98; cited in Hornsby 430)." I adopt this approach because it sees clear to me that social explanations for behavior themselves require explanations, and that these explanations lie in human psychology. (See Tooby's and Cosmides' "The Psychological Foundations of Culture" for extended discussion of this point.)

Second, even if Searle's account can explain how shared social institutions such as language, and its associated notion of proper and improper use of language, can be reduced to individual mental states, Dan Sperber's work warns us that a similar reduction of symbolic action may not be so easy, since we cannot rely on the concept of meaning to effect this reduction. No matter how we understanding "meaning", Sperber has shown us that symbols, in virtue of their evocative power, are far more complex.

Third, although symbols as Sperber understands them have no meaning - that is, there is no unambiguous referent to which they are linked, and therefore not necessarily any unanimity of opinion about their significance, specific symbols do nonetheless serve specific functions within a given society. The Canadian flag, or a state funeral, or a wedding may evoke different emotions and motivations in different people, but there is nonetheless some fixed set of values or ideals that they evoke more strongly and and more widely. In this way, symbolic actions exert their force throughout a community by performing a specific function. And frequently arguments about symbolic utility are arguments about the symbolic functions of certain practices or actions. So to say, for example, that a given action (non-causally) unites, or oppresses, or honors, or offends some individual,

group, or ideology is frequently to ascribe a symbolic function to that action. And I think such claims have an objective truth value; i.e., one that cannot be reduced to simply any subjective opinion of any single individual. That is, when an individual says that a certain practice is offensive, she is doing more than simply reporting her feelings. Her remarks are meant to carry some cognitive content - they are meant to inform us of something about the real nature of this practice. But how is this nature and the function that it enables to be determined when the very function of an action is itself disputed? And is it possible or even philosophically prudent to give an account of symbolic function that is consilient with the ways in which function is ascribed to biological entities and human artifacts? So this is no small problem in the analysis of function and we must settle this matter before coming to a clear understanding of what symbolic function is.

There are two ways we might go here - the first (which Searle adopts) contends that all function ascriptions - including ascriptions of natural function - rely on a prior assignment of value without which they are meaningless. The other prospect (which I defend) suggest that the function of an entity can be identified by its etiology.

Searle's Social Ontology

Searle takes as his starting point the observation that human institutions - money, government, laws, etc. - are just as real and as objective as natural features such as mountains and trees are, in that they exist no matter what we may believe about them. And yet at the same time, institutions exist only because we do believe in them (Construction xi). How is this apparent paradox possible? Searle argues that precisely three elements - collective intentionality, constitutive rules, and the assignment of function - are required to build a social reality from the basic facts of physics and evolution (13).

Collective intentionality occurs when individuals individually form intentions to create something together as a "we." They need not agree to do so, and they need not make this intention explicit. What is important is that each individual intends to perform this activity as a "we" and not as a group of individuals exercising individual intentionality. The intentionality here, while it occurs in individuals' heads, is about group, not individual, intentions. And such intentions, Searle argues, can only be expressed as "we" intentions. They cannot be reduced without remainder to "I" intentions because "I" intentions - even when they are parsed recursively to express the further sentiment that I intend that you intend that I intend, etc. - "do not add up to a sense of collectivity (24)." When, for instance, I say that I intend to steal third base as part of our plan to win a baseball game, my individual intentionality is derived from a collective intentionality. Searle thinks this approach frees us from the false dilemma of being forced to choose between an overly reductionist reliance on "I" intentions and appeals to an ontologically implausible Hegelian world-spirit (23-5). It is this collective intentionality that creates what Searle calls "social facts". Hyenas who hunt together and who coordinate their hunting as a group are not simply

individuals who happen to be hunting the same prey; their collective intentionality establishes the social fact that they are hunting together.

Jennifer Hornsby, however, thinks that Searle's conception does not satisfy Popper's condition. Searle is firmly committed to the notion that all of reality is composed of atomistic particles organized into systems (Construction, 6). Given this, there is no way that Searle can posit a collective intentionality that is irreducible to individual intentionality.

> That which engages in cooperative behavior, when its members each derivatively have an appropriate intention, seems to be irreducibly social. It seems to be constituted (partly) from people's taking themselves to belong to it - from its members each being able to speak of it using "we" .... Must not collectives come before mental states, in order that they can be represented by individual brains, giving rise to "we" intentions? Yet it seems as if collectives could not come before mental states - not if the crucial element in their intentionality is a 'sense of doing ... something together", if "'We consciousness' cannot be reduced to individual intentionality" ([Reality] p.24). (Hornsby, "Collectives" 430, 432)

So, Hornsby contends, Searle is in an ontological pickle: collectives can't be represented in the brain unless they exist, but collectives only exist in virtue of those representations in the first place. But notice that there are at least three ways in which we can understand collective intent :

1. A mental state of a collective when it thinks about itself.
2. A mental state of an individual when she thinks about a group of people that is a group exactly and only because they share some set of

common collective intentions.

3. A mental state of an individual when she thinks about some group of people that the individual considers as a group for some reason other than the reason given in (2).

Searle, I think, is committed only to (3). There is, after all, no incoherence in saying that individuals can have intentions towards several objects considered as a whole ("the junk in my back yard," "those trees across the river," "this handful of jellybeans") in much the same way they can form intentions about individual objects, and that they can do so without our supposing that those objects must exist as a natural kind or that they must share some ineffable quality that enables them to be recognized as a group. All that is necessary for an individual's thought to be about a group is that she considers them a group. And it is perfectly reasonable to say that intentionality with respect to individual items within a given group derives from intentionality about the group - if I intend to burn all the junk in my back yard, I therefore intend to burn each individual piece of junk.

These considerations are just as pertinent when the individual counts herself as a member of the group about which she has collective intentionality, and there are any number of ways in which intentionality about more than one human can arise without us supposing that some illicit appeal to a collective being made ("all of us here now" is one way). As Searle points out, one can have "we" intentionality without anyone else sharing a similar "we" intentionality, and one can even have "we" intentionality when there is no "we" to be found.

I take myself to be engaging in collective behavior with other people, but whether or not I am in fact succeeding in engaging in collective behavior

> with other people is not a matter of the contents of my head. The existence of collective intentionality does not imply the existence of human collectives actually satisfying the content of that intentionality. But once you have collective intentionality then, if it is in fact shared by other people, the result is more than just yourself and other people: collectively you now form a social group ("Responses" 450).

If these comments can be taken to establish that Searle's account can plausibly explain how we can form conceptions of collectives in a non-question begging way, we can then turn to the more central issue of how collective intentionality can create social facts. What is unique to human collective intentionality, says Searle is our ability to

> ...impose functions on phenomena where the function cannot be achieved solely in virtue of physics and chemistry but requires continuous human cooperation in the specific forms of recognition, acceptance, and acknowledgment of a new status to which a function is assigned. This is the beginning of all institutional forms of human culture and it must always have the structure X counts as Y in C ... (Construction 40, emphasis in original)

Institutional facts are then a special subclass of social facts that come into being through the use of constitutive rules that impose status functions on them. Certain pieces of paper, for example, count as money in a given economy only because we ascribe the function of "money" to them. Without this ascription, there is nothing in their intrinsic makeup that allows them to function as they do.

Searle thinks that in fact all functions are ascribed in this observer-relative way. That is, humans never discover functions as intrinsic, ontologically objective features of objects. Rather, they always ascribe functions to objects relative to some value system. Searle contends that some functions are agentive since they are defined by the use to which an agent puts some entity, while non-agentive functions are those that we impose on natural processes (23). And this includes biological functions as well: although we think that we "discover" biological functions, we do so only "within a set of prior assignments of value (including purposes, teleology, and other functions.) (15)" What allows biologists to ascribe functions to organs is that they have already assumed that survival (of the individual or species) is the value by which an organ's function should be judged. But, Searle emphasizes, this choice of value is arbitrary: if one thought that extinction was the endpoint of all living things, one would describe functions very differently indeed.

As I hope to make clear a bit later, Searle's analysis has obvious and serious flaws as an account of biological function. But why worry about this? Our concern here is to explain social, not biological, function. And from this perspective, Searle's position has some *prima facie* appeal. After all, the fact that we ascribe a "money" function to arbitrarily chosen objects is quite clearly dependent on our values, and if those values vary between individuals, they will ascribe different functions (and different measures of functionality) to the same object. Searle's observer-relative account makes this much clear. However, like Searle, I want to offer an account of function that will be univocal across the range of biological entities, human artifacts, and human social institutions (including symbolic actions).

And this is not just for reasons of conceptual neatness. If, as I've argued, the roots of human action (and human interaction) lie in our evolutionary past, the divide between "purely" biological functions and "purely" social functions will prove to be illusory. As Richard Dawkins (in The Extended Phenotype) points out, an organism's phenotype does not end at its skin - it interpenetrates its environment in ways that rely heavily on genetic influences. Second, humans frequently ascribe functions to social entities in ways that impute a status greater than agentive function. That is, when an individual says that the function of free markets is to provide an equilibrium between demand and supply, she need not imply that anyone, including herself, actually uses a free market for this precise purpose and for no other. Searle seems to think that function ascriptions are satisfied whenever a function is assigned relative to anyone's interests. But if this were so, it is difficult to see why people should argue over the function of some entity or other. And it is plainly problematic in the case of a symbolic action that may have a function that is unrealized by many symbolic actors and their observers.

Of course, Searle is free to use the term "function" in any way he wishes. My rejection of Searle's subjectivist definition of function does not imply that there is some realist position which makes it wrong. My usage carries only the pragmatic virtues of making our understanding of symbolic actions and their role within social systems a bit easier and of according (I think) more closely with common usage than Searle's definition does. On the other hand, my definition is not merely stipulative, since the concept of function carries with it a set of normative assumptions, and so therefore it is worth considering as an argumentative definition.

The Function of Function Ascriptions

I will return to Searle in more detail later. But before doing so, it might be helpful to think about just what we want from an account of function.

Functions are (Possible) Effects

First, our analysis should recognize that an object may have no, one, or many functions. Consider what Alvin Plantinga calls the max plan. This is the set of ordered doubles (C, R) comprising all physically possible circumstances and an entity's (determined or probabilistic) response in each circumstance (23). Pretty clearly, not all of these effects will count as functions. The human heart both pumps blood and weighs less than the planet Mars but only the former counts as a function.[47] If every (possible) effect was also a function, describing an entity's functions would entail nothing more (and nothing less!) than enumerating its max plan. And, for some entities, nothing in the max plan will count as its function. Rocks on distant and uninhabited planets, for example, insofar as they have not been the objects of intentional thought, do not have any function.

We should also note that an entity may have some effect F even though the entity is not sufficient or even necessary to bring about F , since many systems rely on redundancies (Wright; Millikan, "Biopsychology"). In fact, the entity may never make F occur (a fire alarm system that is never used, for example), or may not even be able to make F occur (Millikan, "Biopsychology" 212). Wright suggests that ineffectual laws, for example, still have a "function" even though we may

[47] So as not to beg the question against Searle, I should add the proviso that no-one has ever ascribed the function of weighing-less-than-Mars to the human heart.

highlight their ineffectuality by using scarequotes (367).

We should also resist the temptation to identify functions as useful effects. Many functions are useful, and in many cases this explains why they are functions. But many effects are only accidentally useful, and humans occasionally contrive devices whose functions are deliberately useless (Wright, 352).

Functions Occur Within Systems

To say that entity X has function F implies that X does F (or could do F, or is disposed to do F) within some system S.[48] That is, entities, considered as parts of systems, may have functions, but considered as wholes, they do not. To say that the heart's function is to pump blood only makes sense when the heart is situated within a circulatory system, and that system is itself located within some larger organic system.

But, it is objected, suppose some nano-technologist builds a tiny molecular device called the "nanite." The nanite performs one only interesting activity: it flawlessly converts any atomic material it encounters into more nanites who in turn create more nanites until every shred of matter in the entire universe is consumed by and composed of nanites. In such a case, wouldn't we say that the nanite's function is to self-replicate, even if the nanite is not a part of any greater

[48] I define "system" as Peter Munch defines "structure": "A patterned relationship between differentiated parts constituent of a complex whole ("Function," 196). For my purposes here, I shall largely ignore systems that do not arise through either natural selective process (i.e., biological systems) and ones which arise through intentionality (computers, governments, schools, tools, etc.). So this account explicitly excludes, for example, glaciers, weather systems, and planetary systems.

system?[49] (We can imagine that the nanite's effects were not intended or foreseen by its "inventor.")

I don't think so. Just like nanites, earthquakes, floods, or volcanoes just happens to destroy everything in their path, but it is not the case that natural disasters have an all-destroying function. The facts that the nanite is a human artifact, and human artifacts usually have functions, should not lead us astray. While the nanite owes its causal powers to its inventor, its inventor did not, *ex hypothesi*, intend it to have these powers, so those causal powers are, like the causal powers of natural disasters, brute facts to which no function adheres. (For now, I'll avoid the question as to whether human intentionality is either sufficient or necessary for the ascription of function. At this point, I just want to point out that there is no aspect of human intentionality here that provides any indication of the nanite's function.)

Since an entity can be at one and the same time a part of many systems, it follows that an entity may have several functions. So television fulfills one function within a viewer's information-gathering system, another in an corporation's advertising system, and yet another within a media corporation's profit-making system. (And perhaps a further propagandist effect within someone's political system.) These functions interlock: each depends on the successful execution of the others for its own success. And jointly they make the institution of television possible. Proper function ascription depends in part on being able to identify the relevant system within which an entity is located. For example, if people use the shadow of a large rock as way of telling time, a functional explanation of this phenomenon may not explain the rock's geographical position within a given

[49] Randy Wojtowicz suggested this counterexample to the system-relativity of functions, p.c., June 2000.

landscape or its particular shape. But it will explain why the rock continues to play a role in a human timekeeping system.

It also follows from this that it is consistent to say a type of entity has a certain function, whereas a token of this entity may have a function that may or may not be an instantiation of the type's function. Wright suggests that common language usage already recognizes this distinction: "The function of screwdrivers is to screw screws in or out." "This screwdriver is functioning as a doorstop." (353)

Failure to appreciate this distinction can lead to considerable confusion. Igor Primoratz, for example, observes that humans frequently have sex for non-reproductive reasons and that they do so in ways that make conception unlikely.[50] Since, in these cases, reproduction is neither intended or likely, Primoratz concludes that it is not true that the function of sex is to reproduce (16-17).

But this is as bootless as arguing that because some profoundly deaf people use their ears for holding earrings or eyeglasses, the function of ears is therefore not hearing; or because some people use hammers for doorstops, the function of

[50] Primoratz's contention is particularly anachronistic. He cites approvingly Joseph Fletcher's argument that, given uncontroversial facts about menstrual cycles and menopause, women are only fertile for about one day out of any seven between the ages of 14 and 66 (when normal sexuality ends), and that therefore any given sex act is statistically highly unlikely to result in conception. Therefore, he concludes, the purpose of sex is (if these statistics mean anything) to not conceive. But Primoratz ignores the facts that (a) there is absolutely no evidence to show that men mate discriminately with women of any age between 14 and 66, or that they have sex indiscriminately at any time during a women's menstrual cycle, and considerable evidence to show that men are more likely to mate with women who appear younger (and therefore are more likely to be fertile), (b) women in the EEA were more likely to die by the age of 35 or so, rather than over 66, as Fletcher's' example demands, and (c) even if one sexual act is unlikely to produce conception, this in no way implies that a successions of sexual acts could not reliably do so, just as a succession of scratching, wing-flappings, or chewings may reliably perform some function where a single iteration will not. Understanding the function of sex, in other words, requires a closer attention to the entire context in which it evolved, rather than singling out a single instance.

hammers is not to drive nails. The fact that contemporary humans use sexuality (or rationality, or any of a host of other traits) for some purpose today in no way proves that the historical evolutionary forces that shaped that trait were aimed at the same purpose. Nor does the fact that some humans engage in sexual activity on specific occasions for some proximate and non-reproductive purpose prove that sex in general does not (or did not in the EEA (environment of evolutionary adaptation)) achieve some other, quite different, distal purpose.[51] It does not establish that human intentionality is even relevant to determining the function of sex.

The fact that entities may have multiple functions should make us wary of claims that the function of a given institution has been discovered: "The function of pornography is to oppress women." "The function of police forces is to enforce the law." "The function of female genital mutilation is to control women's sexuality." "The function of sociobiology is to legitimate inegalitarian political systems." Whether or not any of those claims are true depends on three further claims:

1. The entity actually does (or is intended to, or could in the proper circumstances) have such an effect,
2. The effect is actually a function[52], and
3. The entity has no other function.

[51] Evolutionary psychologists argue that animals do not seek directly to maximize fitness through their behavior, since the goal of maximizing fitness is too abstract to have been the object of selection. Rather, animals seek sub-goals that tend to increase fitness. In humans, at least, emotions and desires act as the spurs to achieve these sub-goals. The fact that humans also have the capacity to impose additional meanings to their behavior (e.g., engaging sex for hedonistic, aesthetic, religious, or political reasons) does not count against the claim that human sexual activity would have reliably resulted in reproduction in the ancestral environment.

[52] The precise meaning of this clause is, of course, the topic of this section.

Unless someone can provide evidence for all three of these claims - and notice that claim (2) further requires a plausible analysis of what a "social function" is) - any claims that social institutions have a single (political) function are highly suspect. This is especially pertinent in the case of symbolic actions, which frequently rely on shared intentionality (and little else) but where individuals can still disagree meaningfully and strongly about a symbol's import.

Finally, our account of function should preserve the commonsense assumption that one can make erroneous function ascriptions: "The heart's function is to oxygenate the blood." "The car battery's role is to inject a fuel and air mixture into the cylinders."

Natural and Artifact Functions

Traditionally, philosophers have identified two important and apparently distinct types of function: biological function and artifact function. As Elliot Sober (Biology 82) points out, we have a much easier time understanding the functions of human artifacts because we can, in almost every case, infer them from the (assumed or known) intentions of their makers or users. The functions of biological entities are not as easy to determine, and it is harder yet to construct a compact, counterexample-proof explication of function that embraces both biological and artifact functions. Alvin Plantinga thinks that artifact function comprises the paradigmatic sense of function and that all attempts to define "function" naturalistically fail, effectively rendering all functions as artifact functions, some of them supernaturally designed (194-215).

For those who don't take this path, there are only two alternatives: (a) describe natural function in a way that embraces intentional as well as non-intentional function, or (b) (as Searle does) make the functional ascriptions fully intentional, but insist that the intentionality remain observer-relative.

On my reading, it seems that Searle's deepest concern about (a) is that all purely naturalistic attempts to define "function" are doomed because the term "function", and its conceptual cousins, "malfunction", "design", "goal", etc. are all teleological terms and teleological terms are inherently intentional. So defending a univocal account of function will entail first that we explain how teleological terms can be employed within biology in a non-intentional manner.

Charles Darwin's great achievement, claims Searle, was to have driven teleology (and the values it implies) out of biology (16). And thus, he says "... except for those parts of nature that are conscious, nature knows nothing of functions (14)." Biologists and philosophers of biology, however, are sharply divided as to whether Darwin did indeed drive teleology once and for good from biology or whether he actually succeeded in making it respectable. David Hull, among others, agrees with Searle's view. But Elliot Sober argues that, "rather than purge [teleological ideas] from biology, Darwin was able to show how they could be rendered intelligible within a naturalistic framework (Biology 83)."
Michael Ruse also argues that Darwin did not eliminate teleology from biology, and that teleology is not ineliminable from modern biology, but is thriving within it. Ruse points out that even though not all biological features have purposes which were selected for, heuristics that assume that biological features do have purposes (optimality models, for example) pay huge dividends in terms of fruitful predictions and unifying explanations (Darwinian Paradigm, 146-154).

Darwin did show (Searle is partly right here) that there was no "life force," no *elan vital* that propelled evolution and pointed out some goal for all of life, no over-arching *telos*. Darwin famously reminded himself "never to use the words higher and lower [sic] (cited in Mayr, 43)." Life itself is a blind and purposeless force, not in any way directed at producing organisms or species that are part of some larger plan or purpose, but this does not mean it cannot create purposeful and directed organisms. An anonymous critic of Darwin inadvertently captured exactly the import of his theory in what he considered a devastating caricature: "In order to make a perfect machine, it is not requisite to know how to make it (cited in Dennett Idea, 65)." Nonetheless, it seems that the interesting debate here is not about Darwin's actual intentions or his intellectual accomplishment, but over a disagreement as to what one commits oneself to by making a teleological claim ("the heart's function is to pump blood") about some biological entity rather than a merely causal claim ("the heart pumps blood").

Ernst Mayr points out that the respectable use of teleological explanations in biology need not commit the biologist to making four widely-cited errors. First, teleological explanations do not imply the existence of any dubious metaphysical substances such as a "vital spirit" or a "life force." Second, teleology does not make a supernatural end-run around the naturalist dictum that all explanations must be causal explanations appealing to physical laws. Third, teleology does not commit the fallacy of positing backward causation by claiming that a given function exists because it will have some future effect. And finally teleology does not anthropomorphize processes that are not the result of human intentionality. This is because one can intelligently speak of "goal-directed" action without implying that it is the result of any intentionality or consciousness (39-41). And

this "goal" is not observer-relative - one can quite reasonably say that an earthworm's goal is to find food simply by noting its behavior does lead it to food, that the fact that it leads it to food is the reason why the earthworm's behavior persists, and that this is so no matter what anyone's intentions are. In fact, Mayr claims, framing descriptions about biological activities in a way that recognizes that they are aimed at some end has been perhaps the most fecund research heuristic in biology (54-5). So to define "function" in an observer relative fashion, rather than in a teleological one, would be to drain the term of all of its significance within the life sciences.

Take this sentence as an example: "The Wood Thrush migrates in the fall into warmer countries in order to escape the inclemency of the weather and the food shortages of the northern climates (55)." This ascribes a function to the Wood Thrush's behavior, but it does so without ascribing any intentionality to the Thrush whatsoever. But this sentence is not explanatorily equivalent to saying: "The Wood Thrush migrates in the fall into warmer countries and thereby escapes the inclemency of the weather and the food shortages of the northern climates" since it does not explain why the wood thrush flies south. Moreover, to eliminate such terms from biological explanations would also eliminate the fundamental distinction between animate and inanimate objects (56-7).

Still, many object that we should use teleological terms in a biological context only when we do so figuratively: biological talk about directedness and so forth should always be parsed as "as if" talk. For some terms this is clearly so. When we speak of a gene, organ, plant, or animal (excluding the great apes, perhaps) as "trying" or "wanting" or "seeking," this is harmless so long as we remind ourselves that this talk cannot be taken literally.

But it isn't clear that the same is true of all teleological talk. It isn't, for example, true to say that the heart functions "as if" it's pumping blood or that it's "as if" the heart's function were to pump blood. Natural functions aren't "virtual" or "as-if" functions: they are real, in the sense that nature preserves biological entities because of the functions they perform.

So, prima facie, there seems to be no reason to expunge teleological claims from natural science. Nonetheless, on Searle's account, there seems to be no positive reason to include them either, since "nature knows nothing of functions(14)." I'm not certain exactly what Searle means here, but one fairminded way to interpret him is this: of all the possible effects $\{E_1...E_N\}$ of any feature F (where F is pumping blood or making a lub-dub sound, for example) of an object or organism, we cannot pick out any effect E_F as the function of that feature except insofar as it is a function within some agent's value system, and there is no aspect of E_F such that F will have different causal properties or that different lawlike regularities adhere to F. As Daniel Dennett somewhat puckishly puts it,

> It turns out, then, that function talk in biology, like mere as-if intentionality talk, is not really to be taken seriously at all ... Airplane wings are really for flying, but eagles' wings are not. If one biologist says they are adaptations for flying and another says they are merely display racks for decorative feathers, there is no sense in which one biologist is closer to the truth. If, on the other hand, we ask the aeronautical engineers whether the airplane wings they designed are for keeping the plane aloft or for displaying the insignia of the airline, they can tell us a brute fact (Idea, 399).

But this must be false. In biological systems, some effects are favored by natural selection and others are not because of their differential contribution to fitness. Pumping blood contributes to fitness and making a "lub-dub" sound does not, and this is a fact independent of anyone's pro-attitudes towards survival or reproduction. This distinction nicely parallels Searle's distinction between brute facts and social facts (27). Nature can, for example, recognize rivers and mountains, but it cannot recognize social facts such as national borders. The fact that nature can recognize some biological effect F by (statistically) preserving those entities that perform F because they perform F and by eliminating those that do not suggests that the difference between them and other effects is not merely a matter of an observer's values - whether or not those values play any essential role in defining function. And to say this is not merely to make the tautological (but nonetheless important) claim that natural selection preserves those organism that successfully reproduce. Rather, the etiological account of natural function specifies that it is the causal powers of a trait that explain its perpetuation in successive generations.

The central feature of etiological accounts of function is that they distinguish functions from mere effects by showing that a particular function's contribution to fitness played a causal role in shaping the entity and ensuring its continued existence. Function ascriptions, therefore, fulfill an explanatory role by explaining why an entity is as it is (Cummins 1984, 386).

Ruth Millikan, for example, argues that

> ...for an item A to have a function *F* as a "proper function," it is necessary (and close to sufficient) that one of these two conditions should hold. (1) A

> originated as a reproduction" ...of some prior item or items that, due in part to possession of the properties reproduced, have actually performed *F* in the past, and A exists because (causally historically because) of this or these performances. (2) A originated as the product of some prior device that, given its circumstances, had performance of *F* as a proper function and that, under those circumstances, normally causes *F* to be performed by means of producing an item like A ("Defense" 288).

On Millikan's account, this definition (even if is incomplete) is neither an instance of conceptual analysis or a stipulative definition. Millikan argues that to try to effect a conceptual analysis of function is a seriously misguided project, since it is aimed at producing an account of function that allows us to define function in any logically possible situation, and not at creating a theoretical definition that can play an explanatory role.[53]

But this approach is subject to counterexamples. For one, it is possible that some systems could fulfill all of Millikan's criteria, and yet none of its components would count as performing a function.

Mark Bedau, for example, observes that many alluvial clays arrange themselves in crystalline structures that vary from place to place. Subsequent layers of clay will sediment upon these clays by copying the structure of the underlying layers. And since some structures lend themselves to accurate reproduction more than others, there will be differential reproduction rates between different "phenotypes" of clay. Insofar as these as these clays display variation, reproduction, and selection, they therefore fulfill the formal requirements for a

[53] It is of course possible for an entity to perform (different) functions in two different systems. The example of television which I offered earlier shows how this is possible.

historical account of function, but we would not say that any aspect of the clay's structure has a function. So, Bedau says, all natural functional ascription (and the teleological assumptions within which they are embedded) are illicitly importing values into scientific claims where none should be (649).

Contrariwise, Searle argues that it is a consequence of Millikan's theory of proper function that if the human heart came into existence through some nonevolutionary process, it could not, for that very reason, have a function. But this, Searle claims, is wholly at odds with the fact that the heart's function is to pump blood, no what matter its origin may be. (17-18). (And in fact early investigators such as Harvey made meaningful claims about the heart's function even though they had no inkling of evolution or natural selection.) Whether or not this counts against Millikan's project, it cuts at least equally deeply against Searle's own project , since Searle relies quite explicitly on the claim that the heart's function is to pump blood *simpliciter*, and not merely its "function-relative-to-some-value-system."[54]

But neither of these objections need be fatal. Notice that while an analysis of function ought to fit our considered judgments in some central and paradigmatic cases, our pre-reflective judgments about unusual or unexpected instances need not dictate our (non)acceptance of a plausible theory. So even though it may seem unduly mischievous to deny that a heart that popped out of nowhere without cause has no function, there are good reasons to do just that. And it is important to get clear just what we mean when we ascribe a function to an entity

[54] Of course, Searle can object that his ascription of a heart-beating function is merely shorthand - but shorthand for what? Searle seems to imply that his usage here is representative of the "ordinary" use of the term" function" (17). But if this is so, then Searle is equivocating, because people think that the function of the heart is to pump blood *simpliciter*, and they do not think that to say "the function of the heart is to pump blood" is equivalent to saying that "the function of the heart is to pump blood relative-to-some-value-system."

(whether it is a heart or a symbolic action), since to do so carries more weight than Searle apparently thinks.

One reason why we might be inclined to think that a complex object that was not the product of any intentional or natural design process (such as a human heart that appears *ex nihilo*) does have a function is simply because of its complexity. That is, it is almost unimaginably improbable that any complex and functional object could appear that wasn't the product of design, and for which a functional explanation wasn't therefore available.[55]

If, for example, I arrange some Scrabble tiles to spell out the word "LOVE" and leave them on the dining room table for my girlfriend to discover, it is pretty clear that the meaning (however one construes "meaning") of these four letters will in some way rely on the causal processes (some of which are intentional) that caused them to be there. On the other hand, if I drop a bag of Scrabble tiles and the letters accidentally arrange themselves to form "LOVE" then (absent some story about supernatural forces) no story about how the tiles got there will tell us anything about their "meaning" since any meaning we ascribe to the tiles is, as Searle would agree, extrinsic and merely imposed by us.

But now suppose a group of Scrabble factory workers arrive at their workplace one morning to find that an enormous explosion had destroyed the factory and spread thousands of Scrabble tiles across the parking in neatly ordered rows. Inspired (and, withal, having nothing better to do), they carefully transcribe the resulting pattern, which spells out what we might call an accidental and "as-if"

[55] As Elliot Sober points out, William Paley saw this quite clearly. And while much modern philosophical sentiment concurs that David Hume's objections against the analogical aspects of Paley's design argument are decisive against natural theology, Sober points out that Paley's argument, construed as an inference to the best explanation, is, in pre-Darwinian terms, quite solid, and untouched by Hume's position (Biology, 34-6).

novel, one that moves its readers as deeply as anything written by Leo Tolstoy or Thomas Mann. But, just as in the case of the four tiles that fortuitously spell out "LOVE," there can be no question that any element within this novel has a function. Even though there may be passages that appear to constitute examples of foreshadowing, metaphor, allusion, literary irony, etc., they in fact have no such function, just as the letters in the accidental four-tile arrangement "LOVE" have no function. This is not simply the death of the author, but "his" absolute nonexistence.

As with meaning, so too with other sorts of function: complexity confuses. Even if atoms suddenly self-arranged themselves into exact copies of pencils, woodscrews, human toenails, or eyelids, we would not (and should not) ascribe intrinsic functions to them or their parts (even though we are free to co-opt them for our own projects where they can "function as-if" they were their more mundane *doppelgangers*.) The conclusion here is that, appearances aside, we should no more ascribe functions to the parts of the "coincidental human" than we should ascribe meaning to the Scrabble novel.

Searle has two final objections to non-intentional function ascriptions. The first is that historical accounts are vulnerable to counterexamples such as the claim that since viruses cause colds, and since colds are necessary for the existence of viruses, it follows (contrary to ordinary understanding) that the function of colds is to spread viruses (18). Well then, so much the worse for our ordinary understanding. I see no reason not to say that the function of a cold is to spread viruses around just as the function of burrs and pollen is to spread plants around. The fact that we do not see the viruses involved, and are only dimly aware of them, explains why this fact is not obvious to us, but this fact in no way counts against it.

Finally, Searle contends that these accounts cannot explain how we can say a thing malfunctions except by reference to what that thing is supposed to do. If this normative component is ignored, not only is there no way to distinguish (on merely causal grounds) between a functioning heart and a malfunctioning heart, there is also no way to distinguish between a function ("what it's supposed to do") and a mere effect (what it just happens to do) (18-19).

To the contrary, it is perfectly possible to define malfunction in a non-normative fashion and in a way that is consistent with Larry Wright's construal: malfunction is simply the difference between the degree of function an entity actually displays and the degree that explains why it is there. If, for example, natural selection favored adult human hearts that pumped between ten and twelve litres per minute at rest, then any value substantially outside this range is an indication of malfunction. In the case of biological entities, the degree of functional performance that natural selection has favored could perhaps be determined by the average of measurements taken from representative organisms in their usual environment. For artifacts, the designer's intentions will usually dictate how well a design is "supposed" to function, and thereby the

reason why that design was chosen.[56]

The conclusion of this section is that some etiological account is the only way to meaningfully ascribe function to biological entities, human artifacts, and social institutions. So roughly speaking, we may say that *x* has function *F* iff *x*'s doing *F* (or having done *F*, or normally being able to do *F*, or being disposed to do *F*, or doing *F* in certain circumstances) tends to be the best explanation (chosen from a contrast class of all of *x*'s effects) of why those aspects of *x* that do *F* (in the broad sense of "do" outlined above) are the way that they are. In those cases where we are trying to determine the function of a human artifact and we have no idea of

[56] Searle makes one other objection that perhaps does not merit too much discussion. It is this: the proof that function ascriptions are value-laden and therefore intentional-with-a-t is that they are they are also intensional-with-an-s. That is, an intentional term cannot be substituted with some other coextensive term *salva veritate*. So, for example, rowing consists in "exerting pressure on water relative to a fixed fulcrum", but substituting this term for "rowing" changes the truth value of any intentional claim about rowing:

a. I enjoy rowing.
b. I enjoy exerting pressure on water relative to a fixed fulcrum.

The same, Searle says, is true of functional ascriptions:

c. The function of an oar is to row.
d. The function of an oar is to exert pressure on water relative to a fixed fulcrum.

While (b) looks clearly false, (d) only appears false insofar as it appears to an incomplete description of an oar's function. Oddly enough, the same incompleteness appears in uncontroversially non-intentional claims about rowing:

e. The robot rowed slowly.
f. The robot exerted pressure on water relative to a fixed fulcrum slowly.

This does not imply that the penultimate sentence harbors some hidden intentional language. Rather, it signals that "rowing" and "exerting pressure on water relative to a fixed fulcrum" are not in fact coextensive, since many other objects (rudders, daggerboards, many types of valves, canal gates, etc.) also exert pressure on water relative to a fixed fulcrum.

This problem may prove to be endemic to functional statements, since many functional terms are more general than the instantiations we typically talk about. That is, the heart pumps a particular fluid in a particular way, but the verb "pump" embraces many different ways of moving many sorts of fluids and gases. So it will be hard to find truth-preserving substitutes for "pump" in sentences such as "the heart's function is to pump blood." But the these terms will look equally astray when substituted into purely causal (i.e., non-intentional) claims such as "the heart pumps blood."

the designer's intent (the odd relic found in one's attic or during an archaeological excavation, the unfamiliar and undocumented controls on a new piece of machinery) we proceed in much the same manner as I have outlined here.

Note that although I have rejected Searle's account of function as an essentially value-laden term, this does not expunge value from function ascriptions. Instead, it gives it a new role, a role still quite adequate for Searle's purposes, but one that must be constrained by the criteria I have described above. That is, certain pieces of paper count as money because of the values of the people whose collective intentionality deems it money, and it is necessary that people do count these bits of paper as money, but it is not this fact alone that gives money its function. It is that the intentionality of the persons involved is itself sufficient to explain why money exists and fulfills the role it does. This is an example of what Searle calls agentive function, since an agent's intentionality is necessary for a thing to acquire its function (although not quite as he would describe it). These he contrasts with non-agentive functions that "are not assigned on objects to serve particular purposes but are assigned to naturally occurring objects and processes as part of a theoretical account of the phenomena in question (20)." My chief point against Searle here is that the latter class of function ascriptions are not value-laden. The reason why the biologist says that the heart's function to pump blood while the cannibal thinks the heart's function is to serve as dinner is not merely that they have different values. It is that the cannibal's assignment of function relies crucially on human intentionality, while the biologist's assignment of function picks out the feature that cause the heart to continue to exist.

Elliot Sober's view (excepting the generalization in the first sentence) captures much of what I am trying to say here:

> An interesting feature of all [sic] extant philosophical accounts of what the concept of function means is that they are naturalistic. Although the theories vary, they all maintain that functional claims are perfectly compatible with current biological theory. None requires that goal-directed systems possess some immaterial ingredient that orients them toward their appropriate end states. Whatever association teleology may have had with vitalism ... in the past, there is no reason why functional concepts cannot characterize systems that are made of matter and nothing else. ... Selection processes cause some features of objects to be present because they conferred survival and reproductive advantages in the past. This distinction can give meaning to the idea that function ascriptions apply to some characteristics of an object but not others. (Biology, 86)

In sum, my account of function differs from Searle's in that (1) it is historical, rather than subjective, (2) it is causal, and therefore to be discovered, rather than merely ascribed, (3) it takes biological function rather than artifact function as the paradigm by which to construct a concept of function, (4) it disallows claims that any effect can be a function, (5) it allows us to say that an entity functions well or not, irrespective of any observer's interests, 6) it accords with both common usage and common philosophical usage (i.e., it avoids counterintuitive claims such as airplane wings have real functions while bird wings only have as-if functions), (7) it gives an account of teleology which is useful within biologically and also scientifically acceptable, (8) it allows the concept of function to be used in a manner that helps understand how organisms and systems work (i.e., the concept plays an explanatory role), and (9) it does not allow correct ascriptions of

function to be satisfied merely by the existence of one observer who deems that a given entity has a given function (i.e., we can meaningfully say that some function ascriptions are wrong.)

Constitutive and Institutional Rules

Given this amendment to Searle's usage of function, we can now turn our attention to his account of institutional and constitutive rules. Searle distinguishes between brute facts - those facts (such as the fact that Mt. Everest is the highest mountain in the world) that exist independently of language and institutional facts - those facts (such as the fact that Jean Cretien is the Prime Minister of Canada) that rely on human institutions. Institutions are both created and controlled by rules. Searle further distinguishes between regulative rules ("drive on the right hand side of the road") which "regulate antecedently existing activities (28)" and constitutive rules (e.g., the rules of chess that in and of themselves actually create the institution of chess). Human institutions - like language, money, government, chess, and so on - owe their existence to constitutive rules that impose an institutional character on brute facts manifested as pieces of paper, vocalizations, or even thoughts in one's head (28-35).

Searle employs the example of a stone wall built to repel intruders from a tribe's home territory. So long as the wall is high enough and strong enough to do this, it performs its function through sheer physics. But if the wall crumbles to a mere line of stones, it may accomplish the same function if the tribe and potential intruders recognize, and continue to recognize, collectively, the function of the (now much-reduced) wall to delineate a boundary. And this function is what Searle calls a status function, which can be expressed formulaically as "X counts

as Y in C (39-48)." All this entails several other facts about status functions: people need not consciously impose a status function on some object (since they may simply accept that it has a given status without thinking why it has that status), people may be mistaken as to why something has a given status function, status functions can be fraudulently or mistakenly acquired, and, finally, acquiring a status function typically entails also acquiring a linguistic label (48-51).

Status Functions and Language

Implicit in all of this is the fact that status functions are partly constituted by mental representations: a status function simply cannot exist unless someone believes it to be a status function. Further, imposing a status function on something is also and unavoidably a linguistic act, because there is simply no way for anyone to have pre-linguistic thoughts about status functions (64-6). We can, for example, see a man carry a ball over over a line, and we do not need language to do so. But we cannot see him score six points in the same way, since there is simply no way to think pre-linguistically about points because points cannot exist without language or some equivalent form of symbolism. Neither brute facts or pre-linguistic thoughts are sufficient to create institutional facts, since neither is sufficient to create the convention whereby one thing can represent another. Even a pile of stones that represents points will count as a rudimentary linguistic system, given Searle's minimal requirement that linguistic symbols need only "symbolize something beyond themselves, they do so by convention, and they are public (66, emphasis in original)."

Given all this, there will be no clear and sharp divide between the linguistic and the nonlinguistic, or between the institutional and the non-institutional. Non-human animals, for instance, may have a disposition to avoid the territory of other conspecifics, but this fact by itself imposes no deontic status on the boundaries between those territories (71). The reification and institutionalization of norms regarding territories will co-evolve with the advent of symbolization, as individuals increasingly come to interpret some forms of behavior not simply as signs that are non-arbitrarily related to another's individual's likely future behavior, but as symbols that can be (increasingly) consciously and intentionally deployed to communicate one's intentions, and where the relation between symbol and what is symbolized becomes increasingly arbitrary. So an individual's territorial behavior can evolve from a simple disposition to stay within its territory and to keep conspecifics out to the use of aggression, aggression displays, scent markers, artificial boundaries such as walls (as in Searle's example), and finally to the highly formalized, sharply deontic, and shared institutional rules governing international boundaries to which we adhere today.

As Searle recognizes, this raises a special problem for language itself: if language is an institution, and institutions rely on linguistic facts, then - as it seems - our recognition of language as language must rely on some further linguistic markers that denote which behaviors count as language and which don't. Searle avoids this regress by arguing that language is "self-identifying" - unlike other institutions, it doesn't need representations to explains its significance since children, as language learners, are simply socialized to accept language as language (72ff).

This answer may be too short, and linguists such Stephen Pinker have certainly

much room to argue that the facts of language acquisition militate for the positing of an evolved species-specific language acquisition module that cues children to assume (rather than infer or learn) that vocal utterances are important, contentful, and indicative of intentionality. In any case, Searle requires only a minimal level of symbolization for an entity to count as fulfilling this linguistic function in creating social facts.

Notice that symbolic actions typically have two levels of meaning that ought not to be conflated. In the first place we can say, "Such-and-such behavior counts as a funeral in such-and-such conditions" and this will be sufficient to fulfill Searle's requirement that deontic functions cannot be imposed on brute facts sans language. But this imposes only a formal constraint, since it only sets limits as to what sorts of activities may count as a funeral. It does not however shape the emotive content of a funeral, and this is the second, and misplaced, way in which we refer to "meaning." If I attend a funeral, I may intend to express "I mourn your death" but this can hardly count as the "meaning" of a funeral. Such an analysis does not begin to capture the range of emotions that motivate my decision to attend, the emotions that the funeral may evoke while I attend it, and the way my attendance may change me when I leave. The "rules" of attending a funeral do not specify what emotions I should feel, nor how exactly they ought to motivate me. We do not even assume that each and every one of us should come to "express" the very same feelings or what those feelings should be. What is prescribed is the way in which we ought to express ourselves. It is therefore doubly confusing to speak of the "meaning" of a funeral, wedding, or other symbolic activity for two reasons: such actions do not typically stand in any one-to-one relation to any well-defined value or desire, and the emotions that they do evoke may typically be expansive, inchoate, and ineffable. We can thus avoid the problem of positing ineffable meanings if we simply remind ourselves (as

Sperber points out) that symbolic actions do not have "meanings" in the same way that words do, and that seeking the meaning of these actions is not the key to understanding them.

Status Functions: Recursivity, Contestation, and Polemic Force

Moreover, status functions can be assigned recursively. For example, only individuals with the status of Canadian citizen may hold certain offices in certain situations, and only certain officials are empowered to designate individuals as holders of those offices. Status functions are typically like this - they rely on an interlocking, widely understood and endorsed set of other status functions for their continuance.

In many cases, individuals accept status functions without understanding how they have come to acquire that status, or even seeing that their having that status is in any way problematic. Nonetheless, they can tacitly accept and follow the rules that govern a status function, even if they cannot articulate those rules. Searle contends that this point requires an inversion of our usual understanding of rule-following:

> Instead of saying, the person behaves the way he does because he is following the rules of the institution, we should just say, First (the causal level) the person behaves the way he does, because he has a structure that disposes him to behave that way; and second (the functional level), he has come to be disposed to behave that way, because that's the way that conforms to the rules of the institution (144).

Searle invites us to imagine an isolated tribe in which children grew playing baseball in accordance with the rules of the game (which must be the case, since observing these rules is constitutive of baseball). And yet the children never learn the rules as rules - they are simply corrected when they violate one. In such a case, says Searle, these children are neither consciously or unconsciously following the rules of baseball. Nor is this example bizarre, since it mirrors exactly the ways in which we typically learn language by evolving "a set of dispositions that are sensitive to the rule structure (145)."[57]

Martin Bunzl thinks cases of this sort start Searle on a slippery slope that he has not foreseen. Searle, he admits, offers an appealing way to explain how much of our social behavior conforms to institutional conventions without saying either that it is just habit or that we must be following -consciously or unconsciously - some set of rules. But the clear-cut nature of the rules of chess or baseball should not deceive us, since in many (perhaps the majority?) of cases, the meanings of social actions are contested and different segments of a given community may ascribe different meanings to the same event. Bunzl cites the anthropologist Roger Keesing:

> The same "cultural symbols" may have different "meaning", not only for individuals, but for categories of people with structurally opposing perspectives and interest .... different individuals participate in different

[57] Wes Cooper suggests that Searle does think that these children would be following rules. "When he rejects unconscious rule-following, he has in mind the Freud and Chomsky notions of such." Searle says an anthropologist might come up with the rules of baseball by watching the children, but "it does not follow from the accuracy of the anthropological description that the members of this society are consciously or unconsciously following those rules (145)." Given that Searle thinks this case is directly analogous to language acquisition, where he repeatedly says that having behavior that matches the rules is not evidence that we are consciously or unconsciously following rules, it seems more plausible to suggest that Searle is intent on arguing for a sharp distinction between actions that arise due to dispositions that are sensitive to the rules and rule-following itself.

> ways in preserving, transmitting, and changing the religious practices and symbolic structures of the community (578).

Since it is continuity of rule-following that propagates an institution, Bunzl claims that such rules will be hegemonic and accordingly admit of no possibility of change. If our fictive tribe were to alter the rules of baseball, they would still be engaging in an enterprise of collective intentionality, but they would no longer be playing baseball. So, too, it seems, with other institutions. But it is obvious that institutions (especially the symbolic ones that Bunzl is speaking of) do change, and that change is initiated by the participants themselves (576-9). The passage above also intimates, though Bunzl does not make it explicit, that Searle's account may be too lax as well as too stringent: if function ascription is purely observer-relative, then, where observers disagree, there may be no way to ascribe any proper or canonical function to any action.

But why is this an objection? Well, suppose Jews think that the function of a synagogue is to provide a place where co-religionists can meet to worship. But anti-semites might think that in fact the synagogue's function is to act as a headquarters for a Zionist conspiracy aimed world domination. In such a case, an impartial observer would have to conclude that there might be no proper function of a synagogue at all - but this conclusion would turn only on the question of whether or not there existed any anti-semites who in fact held a conspiracy theory about synagogues. But the fact that there was no proper function would not turn on whether or not the anti-semites were correct about

Zionist conspiracies actually occurring within synagogues.[58]

So clearly something must be wrong with Searle's account. While some institutions (marriage, money, government, etc.) rely crucially on broad consensus for their existence, this is not so for other institutions. It is, typically, the intentionality of Jews that defines the function of a synagogue, and the existence or nonexistence of anti-semites who think otherwise is simply irrelevant, since anti-semites simply do not have the requisite causal powers to define the function of a synagogue. In other cases, a function of a social institutions may not be determined by anyone's intentionality. Invisible hand explanations typically maintain that at least part of the reason why some entity exists and continues to exists is that it fulfills some function that the participants may not intend or even be cognizant of. Nozick suggests that invisible hand mechanisms can bring about some pattern P either by filtering out all non-P patterns or behaviors or by bringing systems into equilibrium at P (Anarchy, 20-22). Free markets - no matter their other flaws - bring the conflicting interests of buyers and sellers into equilibrium even though none of the players may intend this, and the fact that free markets do so is not merely one of their effects, but is rather a function of free markets that in part explains why they exist and

[58] So far as I can see, Searle does not provide the caveat that an entity actually has to do F for F to be a function of that entity. He says, for example, that humans have no functions *qua* humans "Unless we think of humans as part of a larger system where there function is, e.g., to serve God (19, emphasis added.) Searle does not say either that such a system has to exist or that God has to exist for us to have a role within it. It is sufficient merely that we believe such a system exists. Contrariwise, if we don't believe it exists, then we have no role in it (even if the system does in fact exist through the will of God).

continue to exist.[59]

Bunzl's point is especially pertinent to my purposes here, since symbolic actions of the sort I discussed in the last chapter will typically have no "meaning" that it could not easily be described by a set of constitutive rules. But this is not fatal to either Searle's project or mine. Even if symbolic actions - weddings, funerals, religious worship, opening ceremonies, etc. - will evoke different emotions and affects in different people, there will nonetheless be a distinctive and more or less shared set of constitutive rules that govern our behavior toward a set of brute facts (birth, death) or institutional facts (marriage, promotion). And they can accomplish this in two ways.

When a person possessing a certain status function utters "I declare you man and wife" this counts, in certain situations, as assigning a certain status function to two individuals with concomitant rights and responsibilities. But this need not count as a symbolic action, since it possesses utility just in virtue of the changed status of the spouses. But note that events like these will typically also be governed by a second set of constitutive rules that also govern our behavior toward those events and that mark the event as a symbolic event. These rules are what separates the mere exchange of marriage vows from a wedding ceremony proper, and an simple interment from a funeral. This second set of rules ("wear black to funerals", "toast the bride", etc.) play no necessary role in accomplishing

[59] In these cases, collective intentionality obviously plays a role in creating free markets, but it is not directed specifically at achieving price equilibria. Nor is there any natural selective process that explains the existence of free markets. So how are we justified in saying that the (or "a") function of a free market is to bring demand and supply into equilibrium? Philip Pettit calls this the "missing mechanism" argument against functional explanations of social phenomena, since these explanations are apparently undermined by the absence of any mechanism which erects and maintains a social entity for its putative function. Pettit's suggestion is that we imagine a process of virtual selection in which, no matter how people hit on the idea of some social institution, it would have been preserved because it performs that function. That is, it is the fact that the institution performs a given function that explains its resilience (Pettit "Functional Explanation.")

either some purely physical act (such as interring a body) or assigning some status function (such as assigning rights of sexual monopoly to spouses). Nonetheless, they make us aware of the importance and relevance of the underlying event by having the power to evoke our responses in appropriate ways. Searle notes that this effect of symbolic actions also frequently serves a propagandistic purpose designed to enforce the validity of the the function ascription to which they adhere:

> Where the institution demands more of its participants than it can extract by force, a great deal of pomp, ceremony, and razzmatazz is used in such a way as to suggest that some thing more is going on than simply acceptance of the formula X counts as Y in C. Armies, courtrooms, and to a lesser extent universities employ ceremonies, insignia, robes, honors, ranks, and even music to encourage continued acceptance of the structure. Jails find these devices less necessary because they have brute force. (118)

Symbolic practices like these are necessary because some function ascriptions will not be accepted by all, and the status function will not obtain unless some minimal level of shared recognition is achieved. (I have suggested in the last chapter why symbolic actions can achieve this role.) Bunzl notes correctly that many such practices are contested, and wonders how people can challenge or change rules if the rules themselves are constitutive of the institutions. The answer to this, I think lies in recognizing that symbolic actions can be contested (and changed) at several different levels, and that radical change to one level need not entail change to another level. Some of these levels are:

1. The constitutive rules that assign status functions to those individuals who themselves assign status functions, including symbolic functions.

Status functions are observer-relative, but not in the overly pluralistic way Searle defines the concept. On my account, the fact that a certain action performs a given function explains why it exists, but there are only certain ways of ascribing function to an action that will allow it to perform that function, and these ways will be defined by constitutive rules that (in some cases) designate certain individuals who are empowered to assign those status functions. So only certain individuals can declare public holidays, confer honors, perform marriages, etc., and only certain sorts of procedures (winning an election, etc.) will confer certain statuses on one. But the rules that determine who those people are may themselves be challenged or changed, and this need not change the institution or its symbolic import. So, for example, one might challenge, eliminate, or even render irrelevant, the status that clergy have to perform marriages. But this need not change the status function of marriage or the symbolic function that weddings perform.

2. The constitutive rules that govern the assignment of status function to some institution. These may be the rules that Bunzl is thinking of, and they may not admit of much amendment without destroying the institution itself. Canadians are, for example currently debating whether "marriage" is an institution that can obtain between two people of the same sex. If it can, can it hold between more than two people at one time? Between a human and a non-human? Could one be married for a very short period time, say, only five minutes?

3. The constitutive rules that confer a symbolic function on some action. Traditionally, North American weddings, at least the Second World War, have been highly ritualized and almost every detail is determined in

advance. Yet none of these details are necessary either to confer marriage status on the couple or to mark the event as a symbolic one. All that is necessary is some means by which spectators' and participants' attention is diverted from the ordinary and everyday to the symbolic. And, as Dan Sperber points out, anything can serve as a symbol.

4. The emotions, desires, memories, values that a symbolic action evokes. As the last chapter makes clear, these will vary from individual to individual: a wedding may represent wedded bliss, a cynical and materialistic swapping of sexual access for material advantage, or the continuance of a family heritage. All of these can be disputed, and individuals can attempt to persuade each other towards some favored interpretation, by making that interpretation more salient, vivid, or valuable.

In the next chapter, I want to explore these suggestions as way of showing how symbolic action is negotiated within communities.

Chapter V

Symbolic Action, the Self, and the State

What is a man,
If his chief good and market of the time
Be but to sleep and feed? A beast, no more.
Sure he that made us with such large discourse,
Looking before and after, gave us not
That capability and god-like reason
To fust in us unused.[60]

On Nozick's account, scientific, pragmatic, epistemic, and moral principles share several similarities. For one, principles group different actions together under a common rubric by identifying relevant features common to all so those actions can be treated in the same way (Rationality, 3). Just as importantly, principles tell us what commonalities are irrelevant and should be ignored. These commonalities are not (in many cases) obvious or self-presenting, nor is it the case that there is only one principle that will adequately describe any set of commonalities. Moreover, since most or all principles are in some way imprecise, inadequate in unexpected ways, conflict with deeply held intuitions, or prove unsatisfactory in some other way, we appeal to various epistemic, rational, practical, and metaethical criteria[61] to choose between contending principles. Principles thus selected allow us to predict, explain, and - in the case of humans - prescribe the actions of other entities in our environment. They also justify actions, transmit probability, and communicate evidence for what we believe to others (5).

[60] Shakespeare, Hamlet. (1601), act 4, sc. 4.

[61] Including empirical adequacy, universalizability, ability to accommodate hypothetical cases, non-indexicality, absence of appeal to proper names, internal consistency, coherence with other principles, explanatory power, fecundity, conservatism, and so on (Rationality, 7).

Principles and Constructing the Self

So, in brief, principles - often unnoticed and unmentioned - guide an agent as she decides what to believe, how to achieve her goals, and how to conduct her affairs with others. Moreover, if she adheres firmly to well-chosen principles, she will be able to overcome temptations that would otherwise thwart her long-term goals, and she will assure others that she is a reliable and trustworthy social player (9-12, 14-21). In this, at least, it is easy to see the role of principles in constructing a life with unity and meaning: without them, an agent would veer wildly from one desire to another, frustrating herself and alienating herself from others.

However, complexities quickly appear. Life, after all, does not permit us to leisurely and completely investigate any and all principles before we choose them. To the contrary, we acquire many of our central principles in childhood, and then frequently only by happenstance. Moreover, our principles are at any time provisional, incomplete, and inconsistent. If we are lucky, we are able to revise them as experience and increasing wisdom directs, but the principles by which we revise are themselves the principles that are most difficult to acquire and the most contested.

I have so far spoken of principles as if they were nonetheless more or less a matter of choice, and as if choosing were only a matter of calling the appropriate principles to mind for consideration. If this were true, then we could indeed scaffold ourselves together out of principles, constrained only by logic and the need to direct and coordinate our principles so as to attain our goals.

Of course, many times we are guided by dispositions of which we are are not (or only dimly) aware, and which probably cannot be described as falling under a

principle (at least one of which are consciously aware) at all. Not only can an agent's behavior be shaped by these sorts of dispositions, she can also adopt and uphold principles for which she has no corresponding disposition to act. She may for example, fervently believe in patriotic principles, but never act on them, and instead on every instance convince herself that some other good outweighs the claim of patriotism. Such a person is not a hypocrite, but she is a person divided - she cannot reconcile her dispositions with her principles, and the reason she cannot do so is in large part because she is unaware that there is any contradiction to reconcile.[62]

What I want to suggest here is that while principles can't be all that constitutes a self, awareness of them is a large part of what constitutes self-knowledge. Self-knowledge - being able to give an account of one's self - is akin to scientific or moral knowledge in that it is achieved and communicated *via* principles. As I have shown, an agent must first know her dispositions if she is to reconcile conflicts between them that would otherwise stymie her plans. But the mere enunciation of a principle that groups together problematic behavior is only the first step towards reconciliation. Overcoming temptation is not usually a matter of an agent's being aware of the existence of the troublesome disposition (she is too aware of it!) but of overcoming it and to thereby render her actions consistent with her principles. Still, she might wonder why she has this particular disposition. Discovering another previously unrealized disposition that fuels the temptation may suggest a way to lessen her compulsion by satisfying that disposition without satisfying the temptation. Or the agent might decide that the "temptation" is in fact her preferred goal, and that she resisted it only because she

[62] Situations like this, wherein an individual understands that there is a reason to do X, but has no motivation or desire to do X, form the meat of the debate over internalist and externalist theories of moral motivation. See David Brink's Moral Realism, 37ff and Michael Smith's The Moral Problem, 60ff for discussion.

did not understand herself well enough. We can, as Nozick suggests, overcome temptation by brute force, by relentlessly following principles directed at other goals. But this is not the only way. We can unify our dispositions in other ways, and we can and should be willing to revise our goals in the light of increased self-knowledge.

Still, Nozick talks as if self-discovery were only a minor aspect of principled action and belief:

> Only rarely do people attempt to predict their own future behavior, usually they just decide what to do. Rather, the person's principles play a role in producing that behavior; he guides his behavior by the principle... the principles are not evidence of how he will behave but devices that help determine what he will (decide to) do. (12)

Several replies can be made here. First, I am not arguing that dispositions completely determine an agent's future actions. So a self-aware agent is not doomed to gloomily repeat her past actions, errors, excesses and all until her death. Second, there may be a few *übermenschen* who are completely immune to dispositions of which they are conscious and who can create new worlds for themselves every day, and therefore never have to wonder about what their future behavior might be, because they simply will what it will be. And there are no doubt others who are simply incapable of projecting themselves into the future, or who are not even aware that such a task is worth doing. *Contra* Nozick, I think most of us fall between these two extremes. Most of us do attempt to predict our behavior in the future (for example, that we will be hungry), and we

choose our goals (being able to feed ourselves) to accommodate these behaviors.[63] One does not simply decide what to do, one chooses a goal, and then decides how to achieve it (in this case, by buying food today). Most of the time, the principle we use to predict our future behavior is simply to assume that, *ceteris paribus*, our future behavior will resemble our present behavior.

Like most *ceteris paribus* clauses, this one is implicit and unnoticed (at least, Nozick does not notice it) until something goes awry. For example, our appraisal of our current dispositions might not recognize all our dispositions, or we might not recognize that some future change (marriage, childbirth, bereavement, etc.) will radically alter our subsequent dispositions. In the latter cases, it is patently futile to attempt to make any plans without some prediction of how one's dispositions might change.

Let me bring out the differences between my position and Nozick's in another way. Nozick says that a person can use principles for either self-creation or self-legislation, as an "external constraint upon the actions of a separate, distinguishable identity." (13) There is, however, less to choose between the two than it might at first appear. The undesirable elements of our psyche are neither so alien that they must be considered separate entities, nor are they so insignificant that the self-designing mind can ignore them. The unifying of the self requires self-reflection, legislation, and adoption of new and desirable dispositions. Nor is there any firm line to be drawn between the principles on which we act consciously and the ones we do not. An agent might firmly resolve herself to follow some principle (unwelcome as it may seem at the moment), but

[63] It may seem contentious to suggest that people make predictions about their future needs. But many food purchases, retirement planning, holiday reservations, etc., seem based on predictions (and not mere assumptions) about one's likely future needs and desires.

with time her consciously following that principle becomes a habit, and then compulsive behavior, and finally perhaps an addiction to which she is enslaved. For similar reasons, there is no clear line between choosing to act regularly in a certain manner under certain circumstances, and discovering that one is doing so. Principles can express (variously) our resolutions, ideals, commitments, preferences, habits, neuroses, and temptations. There are paradigms of each, but - again - no firm line to be drawn between them.

How do principles unite a human self throughout time?[64] I am perhaps less worried about this than some others are. Too close a concern with maintaining a principled connection with the past, or too narrow a resolution to uphold one's present convictions in the future unite a life, but at (what seems to me) the intolerable expense of precluding change. A life can be full and meaningful even if it has radical discontinuities. If a life must have *arche*, or if it is (as some claim) a narrative, this needn't be determined at its outset. Better that we discover our life's purpose - while being true to its mistakes and sins - as we live it than to confine ourselves forever to the naive scope of our youthful ambitions.

We are, as Heidegger points out, self-defining creatures, but we do not and cannot define ourselves wholly by describing what we want to be. We define ourselves also by observing - sometimes with pride, sometimes ruefully, sometimes with shame - what we are and what we have been. We are neither slaves to our past and our dispositions, nor are we free of them. Rather, we interpret them through principles and negotiate new selves who carry this

64 Two Nozickian themes are in tension here: the closest continuer theory and the doctrine of organic unity. For my purposes, it is enough to say that the former is an epistemic principle directed at determining which of several candidates is the true continuing self, and that the latter is concerned with appraising the value of connection and unity between (among other things) past, present, and future selves.

project forward. Our life plans are built in part from the principles we uphold.

Looking More Closely at Symbolic Action

One way of giving expression to these principles is by showing our allegiance to them through symbolic actions. Nozick's analysis of symbolic utility is a powerful conceptual tool because it offers an understanding of human desires and actions that is not available under conventional utilitarian calculus. For example, some people will repeat actions that have no (or negative) apparent utility and that do not even seem designed to secure any desirable outcome.[65] In other cases, agents do justify their action in terms of a desired outcome, but persist even when it is manifestly obvious that the action is not the best way to secure that outcome. But we need not always interpret these cases as ones in which instrumental rationality doesn't apply - it may simply be that the agent is not aware of her real motivations. For example, citizens who insist on closing stores on Sunday may say they want to protect the family, even where Sunday closings do not benefit families in any meaningful way, and even where other

[65] Two examples: In the 1970s, anti-gay activists argued that homosexual teachers should be banned from San Francisco classrooms on the grounds that they were likely to sexually abuse students, even though, as they conceded, homosexual teachers were no more likely than heterosexuals to do so, and this move would therefore do nothing to reduce total sexual abuse. (See the documentary The Life and Times of Harvey Milk). Whether or not this sort of discrimination can be justified by any other arguments, it appears that the symbolic value of allowing homosexuals to teach - irrespective of any other consequences - was (in the eyes of these activists) all by itself intolerable. Another example: citizens in Arizona, Florida, and Maryland, once enthusiastic supporters of "boot camps" for young offenders, are now learning that (much as criminologists had predicted) that the programs are three times as expensive and have a recidivism rate twice as high (80% versus 40%) as conventional prisons (The Sunday Times (24 March 1996), 4). If those citizens continue to support this program despite its obvious lack of success and despite the presence of an cheaper and more efficient alternative, it may well be that they do so because of the symbolic utility expressed by a tough anti-crime policy that "teaches criminals a lesson". My point here isn't simply that an agent's symbolic commitments (including those she may not be fully aware of) can encourage her to ignore pragmatic concerns in a way that is ultimately self-frustrating. It is also that any proposed solution that ignores the symbolic dimensions of the problem will likely be frustrated as well.

measures, no more costly, may be much more effective. So while the agent may consider herself directed at some particular outcome, she may be in fact be engaged in some other symbolic action and, as Nozick points out, this sort of apparent disregard for nonsymbolic utility is evidence that the action has symbolic utility for the agent (but as I've pointed out earlier, this need not always be the case - some actions are just irrational).

Extreme cases of this sort are evidence of neurosis, and Nozick notes that Freudians explain neurotic behavior in terms of its symbolic value (26-7). But less pathological instances of symbolic utility overriding other values are commonplace. For example, it hardly needs to be argued that for many peoples, the pride, dignity, and self-determination associated with national independence has held enormous symbolic utility that overwhelms other considerations.

We often dismiss quixotic actions with no apparent utility as "merely" symbolic, if not downright irrational. But if Nozick is right, symbolic actions can be appraised by the same standards of instrumental rationality by which we appraise other actions. Moreover, since we are, in Nozick's words, "symbolic creatures," it is difficult for us to imagine a life without symbolic meaning. Symbolic actions provide us with a special utility that other actions may not be able to provide. Awarding an honorary degree to a high school drop-out allows him to feel (symbolically) a sense of pride and accomplishment that he would otherwise be unable to attain. Symbolic actions are expressive, and can thus express and strengthen the links between us and our "highest and deepest"

values (Rationality, 30; Life, 286ff).[66] Moreover, the special kind of utility that is imputed via a symbolic action to the agent is constitutive of the unity of a life and of a community, since (for at least some sorts of symbolic action) it unites her present self with past and former selves, with others, and with her values.

Once one has taken Nozick's point that symbolic utility is to be taken seriously, it is all too easy to explain all actions in terms of their symbolic utility. I want to offer two suggestions why we should not do so. First, it simply isn't the case that symbolic utility plays an important role in all our actions. We perform many actions with no consideration for their symbolic import, and would be hard pressed in many cases to explain what the nature of that symbolic utility was. It seems better to restrict our attention to those actions where symbolic utility plays an important role. This does not mean that we need only consider overtly symbolic actions (weddings, ribbon-cuttings, wreath-layings, and the like). As Christ showed us, even a humble meal shared with friends can be imbued with profound symbolic content. Second, we will be especially tempted to ascribe symbolic motivations to actions that appear irrational in terms of conventional utility. But unless we have good reason to believe that the agent was motivated by symbolic considerations, appealing to symbolic utility (like appealing to "unconscious desires" or demon possession to explain otherwise inexplicable behavior) is simply an ad hoc appeal. We have to accept the possibility that some actions, including those that are intrinsically valuable, will not appear rational on

[66] One objection here is that to argue that symbolic actions are chiefly concerned with things so transcendent as "highest and deepest' values thereby situates those values in a realm of human life very far removed from what Pat Churchland has dubbed the "four Fs." My response to this is to suggest in the first instance that Nozick on occasion is given to what Daniel Dennett has called "bombastic redescription." (See The Examined Life for more egregious examples.) The value of which Nozick speaks in such a hallowed way may to turn out to be less ethereal than he thinks. If the Machiavellian Hypothesis is correct, our valuing of symbols and what they represent are still attempts to come to grips with the four Fs but, due to the multiple levels of complexity inherent to human social life, now conducted at second or third remove through elaborate political gestures - whether we are aware of it or not.

any account of instrumental rationality, with or without symbolic utility.

Symbolic Utility and Temporality

As I've argued, symbolic actions impute utility just in virtue of the fact that agents, by performing a symbolic action, evoke within themselves sentiments which frequently relate that action to other events or entities. Frequently these entities are temporally related to the symbolic action. Forward-looking symbolic actions express our determination, intent, or resolve, to carry out some plan of action in the future. Since these actions stand for and symbolize future actions of the same sort, the utility of those actions is imputed back to the agent, increasing the value of the present action and making her more likely to continue to repeat that action and more resistant to temptation (26). As Nozick's example has it, the smoker who wishes to quit is able to refuse a cigarette for the first time because that refusal symbolizes her future refusals and the actual utility of those future refusals, imputed back to that first refusal, outweighs the immediate utility of not refusing. (Of course, we note, this need not be an accurate portrayal of successfully overcoming addiction.)

Backward-looking symbolic actions express an agent's solidarity with her past and her accomplishments. Since symbolic utility imbues her current actions with the value of her past actions, it unites her life into a coherent whole by imbuing her actions through time with a shared resonance. When joint backward-looking symbolic actions are performed, they unite the community within a temporally continuous tradition that educates and or reminds its members of their past accomplishments, commitments, and values.

Other symbolic actions are not so clearly temporally directed. When a

community struggles to rescue a trapped miner, (Nozick's example), it may do so partly because it is expressing its concern to protect all lives. But it is not clear that they do so to prevent future similar tragedies, or to express their resolve to respond to trapped miners in the future, or that they do so out of respect for the past (though their actions may also be motivated by memories of past mining tragedies). Despite the obvious urgency of the rescue efforts, it seems closer to the truth to say that symbolic actions of this sort are "timeless" in that the higher values they point to do not derive their importance from historical or future-regarding considerations.

Symbolic Utility and Politics

Because symbolic meanings are frequently shared meanings, joint symbolic actions are possible. And because symbolic actions can unite us with others (and since unity with others is part of what counts as the good life), it is important that we devise ways to jointly express shared symbolic values. But not all ways of of linking symbolic action to our other-regarding actions will be equally salutary. I think Nozick offers one such suggestion:

> There are a variety of things that an ethical action might symbolically mean to someone: being a member of a kingdom of ends; being an equal source and recognizer of worth and personality; being a rational, disinterested, unselfish person; being caring; living in accordance with nature; responding to what is valuable; recognizing someone else as a creature of God. The utility of these grand things, symbolically expressed and instantiated by the action, becomes incorporated into that action's (symbolic)

> utility. Thus, these symbolic meanings become part of one's reason for acting ethically. Being ethical is among our most effective ways of symbolizing (a connection to) what we value most highly (Rationality, 29-30).

Let me raise a few objections to this. First, as Nozick points out, performing a single moral act does not cause one to become a member of Kant's kingdom of ends (or however one construes the realm of morality), it merely symbolizes it (29). Consider the difference between the evidential import of an action and its symbolic import. The smoker I spoke of earlier rightly construes her first refusal action as having symbolic import. But she would not (and nor would we) consider it evidence that she will continue to abstain.[67] Contrariwise, some actions (such as my regularly picking up the paper each morning) are evidence of a sort that I will continue to do so, but they don't symbolize those repeated actions.

Ethical actions are, I think, closer to the former type. But they differ in an important way. It seems it is much easier, in some cases, to deceive oneself about one's moral failings than it is to deceive oneself about one's weak-willed acceptance of a cigarette. And when we see some of our moral actions as imbued with symbolism, this self-deception becomes easier. Performing one striking and life-defining symbolic action that aligns us with what we take to be our firmly held moral commitments may so capture our attention that it blinds us to the ways in which we fail to uphold those values elsewhere. Civil disobedience to protect the environment, so matter how emotionally important it may to the individual, may be worse than doing nothing if it blinds the activist to her

[67] The smoker, remember, is a hypothetical one, and I've simply posited the fact that she is motivated by purely symbolic considerations.

practical complicity in destroying the environment in other ways.

Second, seeing ethical actions as deeply symbolic may direct our moral attention on how those actions relate to us, since it is to us after all that their utility rebounds. But ethical actions are not primarily about us or for us, but for others. We should not perform ethical actions because they unite us with our highest values, but because they conform with our highest values. It is important that we act ethically, but it is also important that we act, and not merely gesture.

Are there other ways to incorporate symbolic actions into our public lives? As Pranger puts it, "symbolism stands in the center of citizenship's dimension of private attitude (Action, 170)."

> We want our individual lives to express our conceptions of reality (and of responsiveness to that); so too we want the institutions demarcating our lives together to express and saliently recognize our desired mutual relations. Democratic institutions and the liberties coordinate with them are not simply effective means toward controlling the powers of government and directing these toward matters of joint concern; they themselves express and symbolize, in a pointed and official way, our equal human dignity, our autonomy and powers of self-direction (Nozick Life, 286).

The liberal tradition has always understood the importance of individual acts of symbolic self-expression, as ways of disclosing oneself to others and internalizing political values (Action, 145, 170). But these are not sufficient. Part of the meaning of government (over and above its purpose) is, according to

Nozick, to enable citizens to perform symbolic acts that serve as relational ties between them and the state, which cannot be performed effectively by individuals, and which express their solidarity and concern of others (Nozick Life, 288). Symbolic actions can do this because they are economical, that is, they condense complex affective states into discrete actions and thereby release emotional tension. Citizens also internalize values by exposure to passive totemic symbols, such as the flag (Action, 170-2).

The public sphere is a deeply symbolic one. Not only do symbols have a communicative function - they symbolize objects - they also make the appearance of objects possible. Thereby, symbols help us to order and control our environment. But symbols can achieve a life of their own, distorting reality, and, in Whitehead's word, "overwhelming the life of humanity (Action, 150-161)." How exactly is this possible?

Nozick has proposed that symbolic utility be integrated into decision theory. Decision theory can help us decide how to achieve the ends we desire and how to manage conflicting desires so as to maximize utility, but it can't determine which ends we should desire. In what sense, then, can a symbolic meaning not be a "good one" as Nozick suggests (Rationality, 30)? A symbolic action, like any other, can be judged as morally good or otherwise, but in what sense can a meaning be good or not?

Goodman classifies symbols as exemplifying a certain property if they possess that property themselves. Thus a swatch of fabric exemplifies, and therefore symbolizes, all cloth with the same colour-property (Languages, 53). Honoring the tomb of the Unknown Soldier symbolizes honoring all fallen soldiers. But

other connections are not so obvious, and the connections are more open to dispute. The Sikh's turban symbolizes to him his continuation of God's pact with Moses and Aaron, but to others it constitutes an affront to Canadian values.

This example also exemplifies two other reasons why the negotiation of symbolic meaning is so fraught with political risk. First, relatively innocuous actions can symbolize values (such as religious loyalty or national pride) that have almost infinite utility. What would otherwise be a matter of negligible interest (a choice of hat) becomes a test of national will. Second, symbolic meanings are not universal. Heterogeneous societies (like Canada) will share few universally shared symbolic values, and many sets of overlapping meanings. There is plainly the risk of conflict, but within this multiplicity of value, there is also freedom.

> A symbol in a linguistic community is validated by the consensus of the members of that community: its relation to "experience" is less important in the degree that the symbol users employ a system of common counters. (Adams, 209)

Nozick recognizes that symbolic actions can be arenas of bitter dispute. He suggests people may be persuaded to take other actions that do not conflict with either their other nonsymbolic interests or the interests of others. But this proposal makes it sound as if people treat the relation between a symbolic action and what it symbolizes in much the same way they think of the relationship between a means and an end. For example, a rational agent should be indifferent between equally effective means to some desired end. Her wish to buy a certain expensive medication will disappear when she learns that an equally effective but less costly remedy is available. But people do not generally think of symbolic

actions in this way. They are not so much means to ends but their proxies.[68]

Even though the symbolic utility *M* of some act is necessarily agent-relative if it is to motivate the agent, this does not imply that it has no utility to other agents. Symbolic utility, or value, is perhaps most potent when shared between many people. In The Examined Life, Nozick argues that a central (and, in Anarchy, State, and Utopia, unnoticed) role of a democratic government is to uphold and reflect, in a public and official way, certain of the values that its citizens hold most dear and that they consider constitutive of the fullness and meaningfulness of their lives. "[Democratic institutions] express and symbolize, in a pointed and official way, our equal human dignity, our autonomy, and powers of self-direction."

Nozick's attention to symbolic value in The Examined Life and The Nature of Rationality marks a striking point of disengagement from the tenor, purpose, and methodology of Anarchy, though it is not of course the only one he could have chosen. Anarchy makes no mention of either the individual's need to be related to others and to the state or the state's role in promoting solidarity. Much less does it address the question of how all this may be accomplished. The state's role as night watchman (as depicted in Anarchy) lets us sleep without worry, but does not, during our waking hours, provide a focus for the display and celebration of collective values. Part of the intent of The Examined Life is, I take it, to argue why and how states ought to uphold symbolic values. I want to argue here, that in neglecting questions about the origins and heterogeneity of what individuals count as their "deepest and highest" values, Nozick leads us into an unfounded optimism about the solidarity-enhancing and uplifting effect of

[68] For example, an art restorer told me that Mexican peasant women speak of, and treat, religious statuary as if each piece were the saint or deity represented. "Poor Jesus! When will he be better?" p.c., Lourdes Ramos, 1987.

symbolic value.

Nozick does not of course argue that symbolic value is a good that should be pursued above all others in a democratic state. Expressing symbolic value is only one of the state's duties; a duty that needs to be coordinated with the state's duty to better the material lives of its citizens. There is a danger in valuing symbolic utility higher than causal utility in democratic decisions. Demagogues can, and do, mobilize potent symbolic icons surrounding some issues in ways that displace the relevance of questions about pragmatic concerns, and which may deflect attention from other questions of greater (though nonsymbolic) import. For example, the key to understanding the passions excited by the question of Quebec's sovereignty may be to understand that, for Quebecois, sovereignty is not primarily a question of empowering themselves to radically transform their society. (Indeed, sovereigntists have been at pains to argue that government policy, social programs, and even currency will be largely unaffected by independence.) Rather, many Quebecois may simply believe that *only* an independent state can fully express and their cultural and linguistic heritage. The overwhelming symbolic value of independence may have, for some, obscured the consideration that, if an independent Quebec is not economically feasible, this dream is hardly likely to endure. Likewise, though the symbolic value of Martin Luther King Day (for African Americans) and the Equal Rights Amendment (for women) cannot be underestimated, it may well be that pursuing some more pragmatic, and less symbolic goal - such as increasing the minimum wage - would do more to alleviate the plight of disadvantaged groups. This is not to argue that the two aims - symbolic and pragmatic - are incompatible. But there may be times when energies expended on lengthy battles over questions of symbolic value may be better deployed towards more mundane innovations. Further, politicians may exploit the highly charged

symbolic import of questions of otherwise limited national importance - flag-burning, the freeing of a notorious convicted rapist, an opponent's supposed marital infidelity - to detract the public's attention from more complex problems such as health care reform, social justice, etc.

There is no easy resolution to these perplexities. I mention them only to show that while many considerations relevant to democratic decision making (for example, benefits vs disbenefits, probability of success, possible side-effects, relative violation of conflicting rights, etc.) are notoriously difficult to quantify, questions involving symbolic value seem especially resistant to principled and rational resolution.

But exactly which values does Nozick propose the state symbolize? The only examples Nozick cites are equality, autonomy, suffrage, and helping the suffering and needy.(Life, 286-9) To these we might add respect for the rule of law, family stability, education, and some degree of patriotism or nationalism, perhaps. Nozick argues that those who disagree with the state's public endorsement of values necessary to forge bonds of solidarity and relatedness in a society do so because they are lacking the fellow-feeling that *we* feel, and that we can, out of embarrassment for their "unconcern", justifiably act in their name. Let's leave the question of whether the "embarrassment" argument is sufficient to counter the arguments, in Anarchy, State and Utopia and elsewhere, that one is never justified in violating the rights of an innocent other for the greater good. My concern here is Nozick's apparent assumptions that (1) right-thinking people will unproblematically converge on some set of solidarity-enhancing values to be expressed by the state, and (2) that those who object do so because they do not wish the state to express any values whatsoever (except perhaps the libertarian

ones of Anarchy, State, and Utopia.)[69] For the reasons offered above, it is at least plausible to suggest that a democratic state ought to express, for symbolic purposes, the liberal democratic ideals Nozick endorses. But it isn't apparent that the state, in so doing, will have done enough to unite itself to the highest and deepest values of its citizens. For one thing, it is at least plausible to argue that some democratic ideals, cherished as they may be, are merely instrumental goods. Freedom that is never employed to seek one's favored goods is an ideal notion. One does not vote for voting's sake, but to express one's political convictions - whatever they may be. Equality is vacuous until one desires equal access to some good - justice, employment, or whatever. When democratic freedoms are threatened, as they may be in time of war, the state's patriotic gestures symbolize that for now, at least, protecting these values should be the

69 One reason why Nozick does not address the question of convergence of values is his belief that value consists in organic unity. While this isn't the place to fully appraise the doctrine of organic unity, I do want to offer a couple of considerations to indicate why I think Nozick' optimism about convergence is unfounded. Nozick thinks that the way to stop the infinite regress of seeking the meaning of everything (and therefore the meaning of everything) is to insist that meaning is linked in some way to value and that "[v]alue is a matter of the internal unified coherence of a thing. That thing need not be linked with anything else, anything larger, in order to have value."

Now unity, order, and coherence are either a matter of subjective appraisal or they are not. Arguably, if organic unity is unrelated to any other thing, it cannot be related to a subject, and therefore cannot be a subjective value. But then it must inhere objectively in the object as a real property. Let the worries about Platonism that arise here be set aside for the moment. More at least needs to be said about the metaphysical realist nature of organic unity. By what faculty do we perceive it? What sort of thing could it be that has internal unified coherence just in virtue of itself, and unrelated to any other thing? What supernatural suppositions (if any) are necessary to make the idea plausible? Nozick considers that unity in diversity is a measure of organic unity (Philosophical Explanations, 425f). But this seems merely to redefine the project without eliminating the problem: if there are objective measures of coherence, of fitting together well, how can there be any objective measures of disparity or difference?

Contra Nozick, isn't it more plausible to suppose that the degree of unity, or the lack thereof, that one perceives is a function of one's interest, desires, experiences, and knowledge? The battlefield, the Petri dish, and the artist's canvas express chaos to one observer, but order and meaning to another, more knowing, eye. Coherence is not coherence *simpliciter*, but always for someone or something. It is not readily apparent that even the noblest endeavor of humanity - the most perfect state, the fairest and most egalitarian of justice systems - would have any value whatever for non-human beings who did not suffer the frailties peculiar to humanity. See further discussion in Philosophical Explanations.

primary concern of every citizen; the actual exercising of one's freedoms is of secondary importance. In time of war, we may be willing to die for these ideals, but in time of peace they give us little by which to live. We live our lives, express our deepest commitments, by exercising democratic freedoms, but not (at least not primarily) in them.

Nozick recognizes most of this, noting that freedoms are valued not so much for themselves but because they allow the individual "...to engage in pointed and elaborate self-expressive and self-symbolizing activities that further elaborate and develop the person." (Life, 287) Collective expression of values is then continuous with this personal self-expression.

John Rawls has argued that modern democracies must squarely confront the fact that its citizens do not endorse any one reasonable comprehensive doctrine. Rather, modern democrats are composed of a reasonable plurality of more or less comprehensive and rational religious, political, and moral doctrines. Given this, it is unreasonable to expect all citizens of a nation to accept liberal democracy as a comprehensive doctrine, and Rawls now presents his principles of justice as part of a political conception of liberal democracy Rawls Liberalism, xvi-xviii.) Two considerations come to light from this. One reinforces our concern that a state's symbolic recognition of only democratic values will not reflect the deepest concerns of its citizens, since many of them do not endorse anything more than a political conception of liberal democracy. Second, the fact of reasonable pluralism and other modern contingencies presents a crisis of symbolism to modern states. Let me explain.

It is nothing new to suggest that states ought to be aware of the symbolic import

of their actions; Machiavelli saw this clearly, and no government has avoided confronting it (The Prince, 93, 106; The Discourses, 461). However, strong and widespread forces make it likely that the strategies of the past will not be as efficacious in the future. First, modern democratic states are far more heterogeneous than their predecessors, and cultural and religious minorities are more assertive in demanding that their heritage and beliefs be recognized by the state and less willing to be cowed into assimilation. Second, modern democratic states are no longer distant (albeit sometimes malign) entities that touched people's lives only at one or two points (war, taxation, land tenure, etc.) The modern state is expected by many to concern itself with a multitude of concerns that leave hardly any part of our lives untouched, and governments do not have nearly the latitude to avoid culpability that they once had. Third, the pervasiveness and immediacy of mass electronic media radically alters the nature of the relationship between the state and citizens. Electronic media's fragmented presentation of political news (and television's visual impact) facilitates reportage of symbolic events, but is less adept at explaining substantive change or complex political discourse.

The conjunction of these social realities argues for unforeseen and profound crises for the modern democracy. Governments are increasingly less able to meet their citizen's substantive needs and desires (due to deficit and debt) but more able to fill at least some of their symbolic needs (through media manipulation). But a diverse and heterogeneous populace will not be content with nonspecific gestures about the value of diversity. They demand instead their specific cultural, religious, and moral values be respected by the government. Sikhs, aboriginals, and Francophones may have little interest in a government that simply values multiculturalism - they, I take it, want their *particular* values respected.

Perhaps the state will not be able to meet all these sometimes conflicting and incommensurable demands. A bilingual state may never be able to recognize and symbolic Quebec's distinct status, for example. The fragmentation of symbolic value translates directly into the fragmentation of the state.

Chapter VI

Evolutionary Psychology

He does it with more grace, but I do it more natural.[70]

Section A: Rationality as Responsiveness to Reasons

In Chapter II, I noted that Robert Nozick considers that being rational is largely a matter of being responsive to reasons, because beliefs and actions which are responsive to reasons are more likely to reach their respective goals of achieving truth or satisfying desire.

Reasons and Natural Selection

But this, Nozick says, brings us to a curious dilemma. If some reason *r* stands in a relationship to a hypothesis *h* in such a way both that we can recognize *r* as a reason to accept *h* and that *h* is likely to be true when *r* is true, how are we to account for these two apparently very different kinds of relationships? On what Nozick calls the *a priori* view, *r* counts as a reason because it is the sort of thing that a rational creature can apprehend as a reason for *h*. But this leaves us at a loss to explain just why *h* is true when *r* is true. The factual view, on the other hand, emphasizes that *r* is a reason for *h* because it stands in a special factual relation to *h* - but this view can't explain why or how we are reliably able to detect the factual connection between them (Rationality, 107-8).

Nozick suggests that the two views can be neatly combined if we suppose that

[70] Shakespeare, Twelfth Night. (1601), act 2, sc. 3.

there is both a factual recognition and a structural recognition between reasons and hypotheses. The factual connection is the one which transfers support from a reason to the hypothesis, while the structural connection is what is evident to us. Nozick does not appear to think our ability to link reasons and hypotheses is the result of hard-wired dispositions to automatically associate reason and hypothesis or a result of operant conditioning.[71] The real explanation, he thinks, depends on a somewhat deeper and more complex connection:

> Acting upon reasons involves recognizing a connection of structural connection among contents. Such recognition itself might have been useful and selected for. The attribute of a certain factual connection's seeming self-evidently evidential to us might have been selected for and favored because acting upon this factual connection, which does hold, in general enhances fitness. I am not suggesting that it is the capacity to recognize independently existing valid rational connections that is selected for. Rather, there is a factual connection, and there was selection among organisms for that kind of connection seeming valid, for noticing that kind of connection and for such noticing to lead to certain additional beliefs, inferences, and so on. There is selection for recognizing as valid certain kinds of connections that are factual, that is, for them coming to seem to us as more than just factual. (108-9)

This is not of course the only way to account for the connection (or lack of it) between human reason and regularities in the world. Skeptics like Hume denied that we could ever be justified in believing that reason conforms to reality. Kant contended that reason and reality were linked, but that reality itself must

[71] Operant conditioning is the process, espoused by B.F. Skinner, wherein an organism could be taught to respond in a certain way in response to a given input.

conform to reason. An evolutionary psychology such as Nozick's inverts Kant's assumptions by insisting that it is reason which is the dependent variable and that natural selection is the mechanism which ensures that rationality does conform with an external reality - at least well enough to allow us to survive.

Nonetheless, there are problems with this approach, as Nozick notes, and until we have considered the general explanatory worth of evolutionary explanations for psychological traits, Nozick's suggestion may even appear speculative. After all, natural selection selects only those features of an organism which offer an advantage within a relatively stable environment. It is blind to those aspects of the environment which are novel or transient and therefore will generally not preserve traits which are valuable only in some very rare and unpredictable circumstances. If those stable features of the environment should change, there is no guarantee that adaptive traits will continue to confer a benefit on the organism. This is Hume's problem of induction again: because we can never be justified in believing that inductive inferences which have proven reliable in the past will prove reliable in the future, nor can we be certain that inference mechanisms favored by natural selection for the reliability of the inferences which they make will continue to make reliable inferences in the future. Furthermore, it is not clear that natural selection would have favored cognitive faculties which would reliably lead us to true beliefs. Many organisms, after all, do quite well without having any beliefs whatsoever, much less true ones. For organisms such as humans who do form beliefs, there may be no adaptive advantage (and much cost) in having true beliefs where even rare errors of a certain sort (say, believing that "there are no tigers present" when there are) could lead to death, and in these cases we have evolved dispositions to make inferences which are false ("there are tigers present" when there aren't) but which nonetheless enhance our survival. Second, an ability to prefer a more

empirically adequate set of beliefs over another set which is less adequate (Einsteinian over Newtonian physics, for example) may confer no benefit whatsoever, and is thus not the sort of trait which could be preferred by natural selection. And achieving other cognitive powers may involve adaptive leaps so great that they cannot be achieved by successive small steps (109-114). That is, having a certain trait T may confer a benefit on an organism - but T may also impose a cost. Moreover, during some stages of evolving a fully functional version of T, T may confer little or no benefit to the organism, and its cost (in terms of fitness) may indeed be greater. In these cases, some generations of the organism (who are slowly acquiring T) may have less fitness than those who are not evolving toward T. In these cases, natural selective pressures will make it difficult for species to acquire T, since natural selection cannot usually allow for short term losses in fitness for long term gains. For more detail see Richard Dawkins' The Extended Phenotype, 38-41.

But what exactly does it mean to be responsive to reasons? As Nozick tells it, it is a matter of weighing the reasons for and against a belief, of noting which reasons count especially against some candidate belief and which count equally strongly for its contenders, of weighing the reasons for and against the belief's competitors, and of noting what considerations might undercut or reinforce a reason for the belief in a given context (72-3).

Nozick thinks that these conflicting and crosscutting considerations are realized or embodied within the brain as parts of neural nets. For some belief S, a reason R will transmit a positive value to the S-node, while a reason against S will send a negative weight. Undercutters and reinforcers will reduce or amplify the strength of the signal in the connection between R and S (73). Of course, a typical belief need not be shaped by the influence of only one neural net. Many such

units may individually feed forward another net which assigns the final weight. These nets would also embody a set of rules that would appraise various claims for their compliance to various rules of rational belief. And each unit may perform a specific role, evaluating a candidate belief on the degree to which it fulfills one of the traditional virtues of scientific belief - variety of evidence, simplicity, prediction fecundity, prediction precision, avoidance of adhoccery, and so on (77-80).[72]

Nozick's most interesting observation here is that we do not need or perhaps even employ rationality when dealing with many fixed features of our environment. Gravity, for example, is such a ubiquitous and unavoidable aspect of our environment that we frequently do not have to form any beliefs about it: our bodies rely on gravity for many biological processes in ways which we do not even notice (except when gravity is not present), and we typically move our bodies and other objects around without intending to make or conscious exploit any explicit claims about how gravity will affect these actions. We have not, for most practical purposes, had to solve the "problem of gravity" as we would have to solve the problem of non-gravity. The same, thinks Nozick, is true of many philosophical problems - the problems of induction, of other minds, of the existence of an external world, of justifying rationality, and so on. These are all "problems" which our ancestors did not need to solve - at least, in the way that philosophers typically wish to solve them - because they inhere in fixed features of the world, and rationality is a tool which evolved chiefly to deal with unpredictably changing aspects of the environment. That is, humans might have to employ rationality (for example) to make wise choices about the shifting web of political allegiances within their community because these relationships are complex and hard-wired responses would not be adequate to model all possible

[72] See the section on modularity (under "Heritability of Psychological Features" for discussion as to how these neural nets can be incorporated into a modular theory of the human mind.

situations and to select the best response (based on probabilistic calculations of possible causal outcomes). But such a model presupposes that successful humans already suppose that there is an external world, that effects have causes, that some regularities of the past can be used to predict the future, that other human bodies act as intentional beings, and so on.

But, importantly, no humans need have arrived at any of these latter beliefs through rational belief-formation processes. Rather, Nozick contends, humans who just acted as if these assumptions are true (without even articulating them to themselves) thrived where conspecifics who did not comport themselves in conformity with these assumptions did not. And it is because rationality did not evolve to solve these particular problems that they remain among the most intractable of philosophical problems. This, as Alvin Plantinga has forcefully noted, is a devastating weakness of any naturalized epistemology, since we can have no assurance that any inferences about matters unrelated to our survival (including our musings about metaphysics and evolution itself) are reliable (in the sense that the answer we come up with will enhance fitness) much less true. Of course, this is not a problem for the evolutionary epistemologist alone, since any account of human cognition must be compatible with the facts of evolution, and must therefore grapple with its skeptical implications. That is, since natural selection would have favored only those epistemic traits which enhanced survival, and would therefore have favored epistemic traits which led reliably to true beliefs only in those cases where true belief was essential to survival, and since it does not seem necessary to form true beliefs in many cases (on the question of Einsteinian vs. Newtonian physics, for example), the facts of evolution, so well as we understand them, militate fairly strongly against the claim that we could ever have reason to think we could overcome deep metaphysical skepticism (of the sort Descartes proposed). And this reason for

skepticism becomes even stronger when we think about our likely epistemic success in matters such as philosophy and evolutionary biology itself, where epistemic success had no impact on survival whatsoever in the ancestral environment. The only way to avoid these implications of evolutionary theory (and Plantinga is, I think, correct in this) is to advert to a supernatural account of the roots of human rationality (Plantinga, 220ff).

Nonetheless, construing rationality as Nozick does offers a useful heuristic for investigating the particular shape of human rationality and explaining why we have the particular capacities and limitations we do. It also reminds us that what counts as a normative theory of rationality depends to some degree on what sort of a world we live in, and that human rationality is not necessarily optimally suited for many of the tasks for which it used today. Human psychology is not an all-purpose, content-indifferent method of weighing evidence and probabilities. Rather, because of the particular path of human evolution (outlined briefly i n Chapter III, 73f, above), humans possess forms of reasoning that are designed to solve a given set of problems, not infallibly, but at least well enough that they allowed our ancestors to survive and reproduce in the environment of evolutionary adaptedness. We solve some problems well and others very poorly. We are prone to particular sorts of biases and make particular sorts of mistakes. An evolutionary perspective is therefore not only useful, but necessary, for understanding how different aspects of rationality can act to correct some systemic errors of other aspects of rationality, and to understand how it is that culture and society themselves, understood as the manifestations and outgrowths of human psychology, can in turn shape human rationality.

The very notion of evolutionary psychology is, of course, one which is scientifically, conceptually, and politically controversial, and in Section B of this

chapter I'll try to lay out a defensible overview of evolutionary psychology and to answer what I count as the most telling objections against it. In chapter IV, I considered the specific selective forces which would have instilled within us a capacity for, and a disposition towards, symbolic action. Hence, I beg the reader to have faith that the rather long and perhaps non-philosophical excursion into evolutionary psychology in section B will be justified as the theoretical underpinnings for chapter IV's explanation of the evolutionary roots of symbolic action. While many philosophers have explored evolutionary psychology as a means to solve some recalcitrant problems in the philosophy of mind (e.g., the mind-body problem, the nature of mental states, etc.), my concern here is not to explain how mentation is realized in a physical world via an evolutionary process, but to explain how symbolic utility can fit into a conception of human reason and action which is informed by evolutionary theory. Since this field is very large, the sources of evidence numerous and vast, and the arguments complex, I trust that it is acceptable if I at some points refer the reader to other sources for support.

Section B: A Limited Defense of Evolutionary Psychology

Nozick, however, is far more interested in exploring the speculative possibilities of evolutionary psychology than in providing a well-reasoning philosophical and scientific underpinning for his claims. Given the adventurous spirit of TNOR, this is not objectionable. Although some philosophers have embraced evolutionary psychology, I think Nozick perhaps underestimates the degree of suspicion, misunderstanding, and outright hostility that EP evokes in other philosophical quarters. Philosophy is, after all, steeped in a tradition that makes EP claims sound bizarre and incoherent. It is possible that we are not always

aware that the sharp dichotomies so prevalent in the history of Western philosophy may prevent us from fairly appraising new developments in evolutionary psychology (or even from seeing how many of our suppositions are themselves more psychological than philosophical in nature.) Some examples:

- Either the human mind is born replete with innate ideas to be drawn out by the skilled educator, as Plato said, or it is a blank slate on which almost anything can be inscribed, as Locke believed.
- Either our actions are the result of blind, unguided animal instinct or they are the product of intentional and conscious thought (and to blur the distinction between them is just crass anthropomorphism).
- Human dispositions (for action or belief) are either mental (in which case they are the products of our environment) or they are physical (in which case they are the products of evolution).
- Either we are the robotic slaves of our genes or we are free.
- Either human nature is fixed, universal, and unchangeable or it is nonexistent.
- Either existence precedes essence or essence precedes existence.
- Either we are independent and unattached individuals with interests and desires formed prior to entering society (as the doctrine known as "abstract [sic] individualism" holds) or we are constituted by society (Alison Jaggar, cited in Kymlicka, 15).
- Either a human action has a (proximate) social cause or a (distal) genetic cause. And the presence of the former is proof that the latter cannot exist.
- Biology or environment.

- Nature or nurture.[73]

These are clear choices and ones in which it seems that only one option is rationally acceptable. Whether they are explicated enunciated or not, they have not yet disappeared from all realms of philosophical discourse and I do not think we can dismiss them as anachronistic relics of an less enlightened age just yet. One has only to consult the literature critical of EP to see how widely they are held. But evolutionary psychology suggests that these stark choices do not in fact exhaust the metaphysical possibilities for exploring human nature and that these dilemmas are in fact false ones and therefore not only irrelevant but downright unhelpful. Much misunderstanding of evolutionary psychology arises, I believe, from the assumption that any evolutionary claim about human psychology must fall under the scope of the least favored horn in the dichotomies listed above, and can therefore be dismissed as such. But this is not so: the picture is more far complicated than this, and recognizing this fact may require a clear renunciation of any Cartesian divide which licenses one sort of explanation for mental events and another, completely different, sort of explanation for physical events.

To that end, this chapter is largely a work of advocacy. I want here to marshal the best evidence for EP and to answer the most common objections against it. In doing so, I won't be able to answer all objections to evolutionary psychology considered as an entire research program. Nor will I attempt to defend every aspect of current research in evolutionary psychology. As Elliot Sober notes, a large-scale research program such as evolutionary psychology is not falsified

[73] The "nature versus nurture" dichotomy is known as "Galton's Fallacy" after the nineteenth century biologist Francis Galton who famously popularized it, and who may have seen it in Shakespeare's The Tempest. It however originates with Richard Mulcaster thirty years earlier (Harris Nurture, 4). It is still widely embraced - and taught - as a conceptual framework for understanding environmental and genetic influences. My students tell me they are taught this distinction in many social sciences classes and have never been told it is a fallacy.

merely because one of its model fails, and this is a feature it shares with adaptationism as a whole. Evolutionary psychology and adaptationism will be vindicated, if at all, in the long run, when their hypotheses make predictions which are confirmed, and if it avoids offering hypotheses which repeatedly fail (Sober Biology, 128ff, 184). This point is in no way peculiar to evolutionary psychology or to evolutionary theory taken as a whole. Precisely the same point can made with equal force about social explanations for human psychology - in just what way are the claims that "parents have a large influence on their children" or "children learn racism" falsifiable? In any event, falsification is a theory with many shortcomings so it is difficult to see why airy claims of non-falsifiability are thought to be so telling against EP - or against evolutionary theory as a whole, for that matter. [74]

Since my project is only to offer an account of evolutionary psychology which makes it plausible that rationality (and especially the capacity to create, recognize, and employ symbolic utility) is the product of evolutionary forces, I needn't, for example, offer any defense for claims that there may be genetic differences between racial groups which explain differences in measured intelligence. For similar reasons (and for reasons of space), I cannot offer a full defense of current evolutionary theory and so I shall simply address those concerns which seem most pertinent to my own project.

---

[74] Nonetheless, they are still made, and even by people who ought to know better Stephen Jones, the senior editor of the Cambridge Encyclopedia of Human Evolution and the author of the recent Darwin's Ghost recently opined in an interview with Martin Levin that the claim that teenage girls cut themselves in order to secure greater parental attention (since this would increase their fitness) is unfalsifiable, and therefore unscientific. But, in fact, this claim is easily falsifiable: all one has to do is show that a sufficiently large number of girls in a variety of cultures and socioeconomic conditions who cut themselves invariably receive less parental investment than those who don't cut themselves and one has falsified the theory as well as one can be expected to. It should be noted, however, that Jones claimed in the same article that Isaac Newton had "discovered" that earth was round (Globe and Mail, "The Trouble with Darwinian Psychology", April 14, D16).

Instead, I hope merely to demonstrate that evolutionary explanations for human rationality and the behavior it generates are at least as plausible as evolutionary explanations for other aspects of human traits and physical function and for the behavior of non-human species.[75] Those who are suspicious of modern evolutionary theory, the assumptions which underlie it, or the scientific processes which provide its evidential support will simply not be able to reason along with me here and must accept, as their only explanatory option, the theory of special creation. This point is more important than it may appear, because many criticisms directed against EP count with equal force against evolutionary theory as a whole. And when I say "with equal force" it is obvious that I mean with little or no force at all. An EP skeptic might argue, for example, that we cannot justifiably offer any evolutionary explanation for the existence of human aggression unless we can show an actual historical series of aggression-specific human genes, different rates of aggressive behavior, and differential rates in fitness due to these differences. And of course we have no such historical record, nor it is it likely we ever could acquire one. But then of course, the very same is true of non-psychological traits such as disease resistance, digestion, pregnancy, and so on. Yet no-one can possibly doubt that natural selection played a central role in the development of all these features of the human organism.

[75] I'm afraid space does not permit me to offered detailed discussion of research programs in evolutionary psychology which offer positive evidence that it meets the requirements of legitimate scientific research. Instead, I direct the curious reader to a few titles which are representative of the current state of research in this field: Stephen Pinker's How the Mind Works (excellent overview of the current state of cognitive science and evolutionary psychology) and Words and Rules (evolutionary examination of the ways in which humans follow language rules), Judith Rich Harris's The Nurture Assumption (study of the role that peer groups (and not parents!) play in shaping child psychology and the adaptive reasons why this happens), Donald Brown's Human Universals (cross-cultural support for the EP claim for the psychic unity of humanity and for species-typical psychological adaptations), Jerome Barkow's, Leda Cosmides' and John Tooby's The Adapted Mind (collection of multidisciplinary essays outlining evolutionary psychology methodology and its applications within specific research programs), Martin Daly's and Margo Wilson's Homicide (evolutionary explanations for human killing employing extensive cross-cultural data), and Peter Carruthers' and Andrew Chamberlain's Evolution and the Human Mind (up-to-date discussion of competing modularity theories and their application).

Even the most cursory inspection of functional biological complexity - psychological complexity or otherwise - should convince one that organisms could not have appeared through a mere chance concatenation of molecules or even a succession of chance occurrences. It is simply too staggeringly unlikely that any deeply complex, adaptive, and functional structure should have arisen by mere chance , unaided by natural selection. This noted, the only available explanations for biological complexity are some "top-down" (supernatural) cause or a "bottom-up" process wherein natural selection creates increasingly more complex organisms. In saying this, I am not denying that many features are not specifically selected for, nor am I denying the relevance of environmental forces. Researchers hardly agree about just how various evolutionary forces work, or of their relative importance, but it is clear that some evolutionary account is the only plausible naturalistic explanation for the existence, variation, and complexity of terrestrial biota. Natural selection is not the exclusive cause of biological functional complexity, but, unless we appeal to supernatural forces, it is an unavoidably necessary one.

Evolutionary theory unifies the biological sciences by explaining how species appeared, evolved, and reproduced and how functional complexity (reflected in the design of an organism's physical and behavioral attributes) can arise via the algorithmic assembly of relatively simple components. Let us consider evolution as a general design process whereby organisms change over time. This process is "algorithmic" in that it is (1) mindless, (2) substrate-neutral (does not depend on a particular architecture to perform its functions), and (3) that, given appropriate initial conditions, it tends toward the same result - the production of organisms with increased (or optimal or maximal) fitness in a given environment (Dennett, Idea 56-60).

The chief, though by no means exclusive, engine of evolutionary change is natural selection. If self-reproducing organisms possess differing genes, then natural selection will favor those genes (or polygenes) which express themselves phenotypically in ways that are more likely (compared to other competing genes) to ensure - reliably, but not invariably - the gene's survival and reproduction in a given environment.[76] Richard Dawkins has described this view as the "selfish gene" theory. Properly understood, I think this concept is an innocuous one, but it has nonetheless attracted far more controversy and misunderstanding than it deserves. Mary Midgley, for example, tartly observes that, "Genes cannot be selfish or unselfish, any more than atoms can be jealous, elephants abstract or biscuits teleological (Midgley, "Gene Juggling"). As Midgley sees it,

> [Dawkins'] central point is that the emotional nature of man is exclusively self-interested, and he argues this by claiming that all emotional nature is so. Since the emotional nature of animals clearly is not exclusively self-interested, nor based on any long-term calculation at all, he resorts to arguing from speculations about the emotional nature of genes...

But this is patently false. Dawkins' selfish gene theory in no way makes any illicit claims about any putative gene-level psychology, nor does it need to. Dawkins is exceedingly explicit that his use of "selfish" here is exactly analogous to biologists' use of the term "altruistic." Here "altruistic" simply describes an entity (either an individual or a gene) which "has the effect (not purpose) of promoting the welfare of another entity, at the expense of its own welfare

[76] By distinguishing between "reliably" and "invariably" here, I am only noting that some forces will thwart natural selection. An environmental disaster such as a flood may randomly destroy some fitter individuals while sparing less fit ones.

(Dawkins, Extended Phenotype, 284)." And "selfish" denotes exactly the opposite. This definition is completely and deliberately behavioristic, and does not import any unacceptable reliance on a selfish homunculus lurking within the gene, as Midgley wildly misunderstands. "As-if" selfishness here is explanatorily adequate and quite defensible. Nor does the selfish gene theory in anyway imply that selfish genes create selfish organisms or that in every case a single gene is responsible for any given trait, as is commonly assumed.

Nonetheless, the charge that evolutionists have unduly imported intentional notions into nature is hardly a new one. As Dawkins sees it, the very roots of the selfish gene theory lay in early criticism of Darwin's adoption of the term "natural selection." To modern ears, this phrase is noncontentious, but many of Darwin's contemporaries charged that nature could not in fact "select" anything and that Darwin was personifying nature. At Wallace's suggestion, Darwin eventually adopted Spencer's phrase "survival of the fittest" (Dawkins, Extended Phenotype, 179-80). But the concept of fitness has proved at least as confusing as the concept of natural selection and Dawkin's suggestion of the selfish gene is simply his way of making the point that the gene is the unit of selection and "fitness" means genetic fitness. The second edition of his Selfish Gene and his "In Defense of Selfish Genes" offer sustained replies to critics of the selfish gene theory.

If the genes which code for heritable traits can themselves change over time (through random mutation, say), then organisms are capable of of large, unpredictable, and open-ended changes (Kelley, 283-311). But natural selection is, in Richard Dawkin's words, a "blind watchmaker" - wasteful, unpredictable, non-teleological, and unmindful of long-term consequences and eventualities. And many of its products are therefore doomed to failure. Genes, moreover, do

not simply build or control organisms by themselves. Rather, each organism is the product of a long and complex interaction between its genes and the forces of environmental change, sexual selection, genetic drift, and the complex and evolving behavior of other conspecifics and members of other species. We now need to examine this interaction in more detail to understand its implications for human psychology.

The Pieces of the Puzzle

Given the explanatory success and theoretical centrality of evolutionary theory to our understanding of biology, it is more than plausible to argue that evolutionary psychology (EP) can also illuminate our understanding of human psychology. As I've noted earlier, Nozick assumes it unproblematically does explain many features of human rationality, and accordingly offers no defense of EP. I am not convinced that that many in the humanities even recognize the potential explanatory value of EP. In any event, to adequately defend evolutionary psychology, it is first necessary to show that (1) psychological features vary between individuals, (2) that differences in psychology will differentially affect fitness, and (3) that psychological features are heritable.[77] I proceed as follows.

1. Psychological Variation. Aspects of human psychology vary between individuals. This is hardly controversial. It is a fact of universal human

[77] It bears repeating that to say that a trait is heritable does not in any way imply that environmental factors ("nurture" for example) cannot affect its phenotypical expression. People understand perfectly that when one says one has inherited one's hair color from one's mother, one is not denying that hair dye and exposure to sun can also affect hair color. But same point, when made a bout a psychological trait, is frequently confused with a biological determinist position which denies the influence of environment.

experience that individuals do in fact vary in their capacities or propensities for intellectual achievement, sociality, humor, artistic expression, violence, child-rearing, personal ambition, etc.

2. Differential Effects on Fitness. The very concept of fitness is a particularly complex and controversial one. As I noted above, Charles Darwin originally accepted the term at the prodding of Wallace to avoid objections directed at Darwin's concept of natural selection, but the term has been reinterpreted many times since then. Early definitions used "fitness" as a way of describing individuals that had specific features (strength, good eyesight, etc.) that would enhance survival and reproduction. Population geneticists, on the other hand, define fitness operationally as the selection for or against a given genotype at a given locus. Ethologists and ecologists use fitness to refer to an individual organism's success at surviving and reproducing (Dawkins Extended Phenotype, 179-84). Elliot Sober points out that fertility selection works from adult to zygote, while viability or survival selection works on the organism from zygote to adulthood (Biology, 57-9). W. J. Hamilton argued in 1964 that the individual was not in fact the unit of natural selection, and that what was being preserved by natural selection was genes, not individuals. Since this was so, natural selection could be expected to favor genes that would ensure their survival and reproduction not only through descendants of the individual, but through the survival of other individuals who contained copies of the same genes. So a gene that disposes an individual to help its siblings thereby helps copies of that gene in those siblings to survive. Hence fitness had to be redefined as inclusive fitness. "Inclusive fitness is calculated from an individual's own reproductive success plus his effects on the reproductive success of his relatives, each one weighted by the appropriate coefficient of relatedness (Dawkins Extended Phenotype, 186,

emphasis in original)." But Richard Dawkins objects that the concept of inclusive fitness is a last-ditch attempt to preserve the idea that the individual (or group or trait, as others have held), rather than the gene, is the unit of selection. And this, he says, is a serious mistake, since in cases of individuals which help relatives, it is tempting to think that natural selection works so as to make individuals "care" about the replication of individuals with genomes similar to their own. (And therefore to think that each gene acts as if it "cares" about its equivalent copy in that other genome.). But this is not so. "It is better to assume that only genes 'for caring' care, and they only care about copies of themselves (191)."

A further complication is that fitness is only a theoretical property, and that evolutionists often measure fitness as it would occur under ideal conditions, which in the real world, never occur: due to nonselective forces which are ineradicably present, actual survival and reproduction rates need not be identical with theoretical fitness (Sober Biology, 58). Sober's definition of fitness is, "Trait X is fitter than trait Y if and only if X has a higher probability of survival and/or a greater expectation of reproductive success than Y (Biology, 70)." *Mutatis mutandis* for genes, this is the definition I have assumed throughout this work. In many places, I have spoken of the fitness of individuals, but it should be understood this simply a shorthand for genetic fitness, and that increasing individual fitness is only a way to increase genetic fitness. Since none of this work discusses kin selection and the implications it has for inclusive fitness, I do not think this minor deflection from theoretical purity undermines any of my claims about symbolic utility.

A second question arise as to whether natural selection maximizes or optimizes fitness or merely satisfices fitness to some lower level of adequacy. Dawkins rejects both approaches. Optimizing ignores the many constraints on

perfectibility that exists in a given adaptive landscape.[78] And satisficing, he thinks, does not do well enough. "The trouble with satisficing as a concept is that it completely leaves out the competitive element which is fundamental to all life. In Gore Vidal's words: 'It is not enough to succeed. Others must fail (Extended Phenotype, 45-6).'" Dawkins' preferred interpretation is that natural selection meliorates - organisms do not simply scrape along, nor do they achieve perfection. Rather, they simply tend to improve over time.

Having settled these technical questions about fitness, we can now see what effect psychology has on fitness. Pretty clearly, some differences in human psychology would have contributed directly to differential finesses within the EEA. Again, this is uncontroversial. Since human psychology drives human behavior, humans who were psychologically indisposed or unable to ensure their own survival, effectively equilibrate their behavior with the behavior of other humans, engage in heterosexual mating, protect and nurture their offspring, etc., would have, on average, left fewer, and less successful, offspring than did humans who were psychologically capable and willing to perform these activities. Of course, none of this implies that all humans would express these traits in the same way, or that there is any single adaptively best strategy to do so, or that any preferred set of strategies will be optimal in all environments, or that that there is any generalized disposition to increase one's own fitness, or that any adaptively favored trait is morally preferable to any other.

3. Heritability of Psychological Features. Some psychological features are heritable - in precisely the same way that some physical ones are. If, as most contemporary philosophers of mind think, mind-body dualism is false, then the

[78] Sewell Wright's metaphor for fitness wherein an organism increases fitness as it travels up hills o mountains of fitness.

nature of mental features will rely on and be explicable by the physical features of the brain - for there is no nonphysical way in which they could arise. That is, since many mental features will either be (as the most plausible current accounts hold) identical with neural states, reducible to neural states, or will be functions of neural states (and perhaps of other substrata). No matter which account turns out to be true, they all concur in arguing that at least some changes in physical states are responsible for changes in some mental states. And since some of the differences between brains are due to differential genetic effects, this plainly implies that genes will account for some differences between mental states of different individuals also. This is plainly the most controversial of the premises supporting EP. It is nowadays completely unobjectionable to assert that a non-psychological trait (such as a propensity to disease, or taller than average height, for example) could be heritable, but to make a similar suggestion about a psychological trait is frequently met with utter incredulity, an a demand for evidence, and subsequent skepticism of that evidence. This, I believe, bespeaks the deep (but often unnoticed) commitment to dualism which many (even sophisticated) observers still retain. Nonetheless, my claim that psychological traits are heritable is strongly supported by four independent lines of reasoning:

First, personality studies of monozygotic and dizygotic twins reared together and apart indicate that about 50% of the variance in personality can be attributed to heredity (Tellegen et al; Bouchard and McGue; Lykken; Harris, "Environment"; Harris, Nurture)[79] . Earlier socialization studies were frequently flawed since they assumed that all correlations between parental and child behavior mst be due to a direct parent-to-child effect. But these similarities could be due to some other factor - notably, a shared genetic influence, and parent-

[79] I have followed Judith Rich Harris's very accessible account of this issue in The Nurture Assumption in this section.

child studies cannot control for these effects. Monozygotic twins (whose genes are identical) and dizygotic twins (who share only 50% identical genes) make ideal test cases for researchers to distinguish between the effects of environment (typically the home environment) and genes. The Minnesota Study of Twins Raised Apart therefore located monozygotic and dizygotic twins who had been raised in separate homes from a very early age. By subjecting these twins and as well as twins raised together to a large battery of standard psychological tests, Bouchard et al have been able to show that monozygotic twins are strikingly similar to each other (and more similar to each other than pairs of dizygotic twins of strangers are) even if they are raised in different homes and have no contact with each other (Harris Nurture, 21-3, 28-32) . One example shows how remarkable these similarities can be. Two monozygotic twins (both named Jim) reunited in adulthood found that "both bit their nails, drove Chevrolets, smoked Salems, and drank Miller Lite; they named their sons James Alan and James Allan (Harris Nurture, 33). This does not of course imply that all similarities between twins are due to the direct effects of genes. Since identical twins have similar appearances and behaviors, these factors will influence the actions of people around them and will tend to make their environments more alike than they might otherwise be. (For example, parents are more attentive to babies which independent observers rate as appealing ("cute") than they are infants which are rated as homely (Langlois, Ritter, Casey and Swain; cited in Harris Nurture, 29).) So the similarities observed between monozygotic twins are due to both direct and indirect effects of genes and it may be impossible in practice to disentangle them (Harris Nurture, 28-9). On this model, it is no surprise that the children of well-adjusted parents tend to be well-adjusted (and vice versa), since they share common genes. The twins studies show that many similarities persist even where individuals are raised by different sets of parents (Harris Nurture,

34f) and hence could not be due to parental influence. It seems to follow form this that monozygotic twins (who are genetically identical) who are raised in the same (and therefore identical) home should be very alike. But this turns out to be false. Monozygotic twins raised together are no more alike than monozygotic twins raised apart. And other siblings raised together are no more similar than siblings raised together. As Eleanor Maccoby and John Martin reported in 1983, the implications of all this were either that "... parental behavior must have no effect, or that the only effective aspects of parenting must vary greatly from one child to the other within the same family (cited in Harris Nurture, 38)." Drawing on Ernst and Angst's study of birth order and personality and her own meta-analysis of birth-order studies, Judith Rich Harris argues that birth order studies show no consistent pattern, and that the latter disjunct is therefore effectively ruled out. Thus, parental influence on children has little or no lasting effect.[80]

Second, many aspects of human behavior ("human universals") remain constant even where there are vast differences in culture or environment (D. Brown Human Universals; Daly and Wilson Homicide). This was observed quite early

[80] This latter claim deserves some closer consideration, since it seems to contradict the notion, popular since Freud, that there are strong correlations between parent and child behaviors and that parental behavior does indeed have a direct and long-lasting effect on child behavior. The eminent psychoanalyst Bruno Bettelheim, for example, observed that autistic children frequently have emotionally distant mothers ("icebox mothers") and inferred (incorrectly, as it happened) that the mothers' coldness caused their children's autism (Harris, Nurture, 27) .

Harris, however, has recently and forcefully argued that this model may be incorrect. She charges that many studies of child socialization are flawed in that they fail to separate genetic from environmental effects (which error the twins studies correct) and that they conflate correlation with cause. Harris points out that correlations between adult and child behavior can be explained in at least four ways: (1) shared genetic influence, (2) parent to child effect, (3) child to parent effect, and (4) indirect socialization from parent to child via the child's peer group. But, Harris charges, many researchers make what she calls "the nurture assumption" - the unsupported presupposition that parent-to-child effects are the major, if not only, cause of parent-child behavioral similarities. Against this, Harris contends compellingly that it is the influence of the child's social group(s) which in fact explains most of the non-genetic psychological difference between individuals ("Socialization" *passim*; Nurture *passim*).

by Franz Boas who noted that

> We find not only emotion, intellect and will power of man alike everywhere, but also similarities in thought and action among the most diverse peoples. These similarities are ... detailed, ... far reaching, ... vast, and related to many subjects. (cited in D. Brown 56)

Nonetheless, many later anthropologists (notably Ruth Benedict and Margaret Mead) emphasized, not the commonality of human dispositions, but the vast differences between cultures. Despite the fact that *Homo*'s early existence (that is, 99% of our career) was spent in one relatively uniform ecological niche (that is, hunting and gathering on the plains of Africa) and the fact that there is relatively little genetic diversity between human groups (since our last common female ancestor lived only 140, 000 to 280,000 years ago (Cann et al, "Mitochondrial DNA," 31-6) and our last common male ancestor lived only 50,000 years ago (Underhill et al, "Y Chromosome," 358-61)), many observers believed that humans had no interesting inherited nature except perhaps a general ability to learn or adapt or some other "highly unspecialized and undirected drives" (Berger and Luckmann, 48). Therefore, it was held, humans were capable of organizing themselves into almost any imaginable social system with equally diverse belief systems.

This view was powerfully reinforced by several theoretically diverse positions. Freud contended that parental influence on children was so strong that, in some cases, it shaped an individual's entire life, encouraging many to think that these influences (rather than genetic ones) were a large, if not overwhelmingly dominant, determinant in shaping adult personality. Social determinists such as Peter Berger and Thomas Luckmann argued that, even though the "biological

substratum" set some minor limits to human plasticity, ultimately it was society that determined every aspect of human nature (49). For example, some linguists believed that language shaped the very limits of human thought. On this account, if a culture did not contain a word corresponding to, say, *schadenfreude*, members of that culture would be incapable of feeling - or even understanding - the emotion which it named (Pinker 366-7). Another example also illustrates this point. Since the color spectrum is continuous, with no obvious "natural" divisions, different societies could (and, some claimed, did) divide the spectrum arbitrarily in very different ways (D. Brown, 11-14). Moreover, it is possible that other cultures could devise world views incommensurable with their own and which simply did not employ the same categories of thought that Europeans considered not only universal but somehow natural and necessary for even comprehending the external world. Edmund Whorf, for example, argued that the Hopis had no concept of time - or at least a concept of time radically different to our own - <u>because</u>, as he believed, their language contained no temporal terms whatsoever (D. Brown, 27-31).

By the eighties, some theorists were arguing that there were no human universals whatever. Rose et al are unwilling to admit that there are <u>any</u> true human universals except a capacity for language and a few physical commonalities, such as being being between one and two metres in height at adulthood and being unable to fly (13-14, 243ff).[81] The BSSRS Sociobiology Group contends that the evidence for human universals (including even incest avoidance) is so sparse that it is "hard even to accept that there are universal behaviors, never mind to discuss whether or not such behaviors are under

[81] On this view, how any human society can be viable without successfully and repeatedly performing those behaviors that seem <u>minimally</u> necessary for human existence and survival - reproductive sex, acquisition of food, avoidance of harm, infant care, conflict resolution, etc. - is vastly unclear. And those who deny the existence of human universals offer no evidence for the existence of alternatives to these behaviors.

genetic control (119)."

I think, however, that this need not be the final word on the matter. Donald Brown argues that while human cultures definitely do differ, there are nonetheless deeper resemblances between them which underlie these apparent differences. Moreover, many celebrated cases of apparently vast differences in cultural beliefs have proven to be illusory or much less compelling than they once seemed.

Brown distinguishes between true universals (those aspects of human behavior which are found in every culture) and near universals (those which occur in almost (say, over 95%) of all cultures). Brown concedes that there may be insufficient anthropological data to demonstrate that a putative universal is in fact present in every known culture. Nonetheless, he contends, if the feature is found in cultures which are geographically distant and not related to each other, and the anthropological record presents no evidence of any culture in which it does not appear, a strong case can then be made that the feature is in fact a universal. Many near-universals, he argues, may in fact be true universals, but simply fail to manifest themselves in a given culture because they are overridden by some other human disposition. [82] Or (as he argues in the case of competitive games) the anthropological record itself may be suspect.[83] Universal and near-

[82] I do not intend to invoke any heavy metaphysical or psychological conceptual apparatus wherever I use the terms "disposition" and "psychological disposition". Following Penelope Mackie, I define a disposition as as "a capacity, tendency, potentiality or 'power' to act .. in a certain way ("Dispositions" 203)." So one might have a disposition to believe, to desire, to certain emotions, etc. Nor do I make any claims that all mental activity can be reduced to dispositions or that dispositions themselves cannot be further reduced. I use them simply as a useful way of positing an intermediary entities between genes which shape psychology and the forms of human behavior they generate.

[83] This move on Brown's part does not render either the theory of human universals or EP as a whole unfalsifiable. See discussion of Elliot Sober's remarks on optimality models in the section on determinism.

universals of this type then can be considered an unconditional or categorical set of human human behaviors.

In contrast to these, conditional universals are those which appear if some other condition obtains, and where the antecedent condition does not describe a culture or another universal. For example, not all cultures value the right hand over the left hand. But cultures which do assign value to one hand or the other invariably assign value to the right (89-90). So, on this account, many cultural differences are not the arbitrary result of environmental noise and historical happenstance, but are in fact the manifestations of conditional universals (41-6). Brown vividly displays the wide range of human universals by describing what he calls the Universal People, who bear those features which all humans bear - abstract, symbolic, and evolving language, facial expressions, toolmaking, use of shelter, incest avoidance, status divisions, male political dominance, conflict resolution, child socialization, division of labor, music, and so on (130ff). Brown, in fact, lists dozens of human universals, the claims for which are in some cases controversial, and it is obviously beyond the scope of this work (or my anthropological abilities) to defend each and every one of them. Nonetheless, if these patterns (some of them surprisingly specific) do permeate all human cultures, there must be some persuasive, non-coincidental, explanation for their widespread existence. Doubtless, some universals are forced on us by the universal conditions of the physical world, while others may have radiated throughout all cultures. But Brown argues that many universals such as nepotism, incest avoidance, and so on which contribute directly to fitness are the products, directly or more remotely, of natural selection (86).

Third, many psychological features (for example, vision, highly adaptable intelligence, and a capacity for language) are complex and adaptively specialized

functions which our very early (that is, prehuman or even pre-mammalian) ancestors obviously did not possess. Since natural selection is the *only* available scientific explanation for organized functional biological complexity, these features must therefore be the products of natural selection. But for any psychological or physical characteristic of a phenotype to be the object of natural selection, it must be heritable and therefore under some degree of genetic control (Tooby and Cosmides, Dawkins Extended Phenotype, 26, Pinker and Bloom).

Fourth, it is not a reasonable hypothesis that humans could possibly solve the multifarious problems required for them to survive and reproduce without the benefit of a sizable complement of "preprogrammed" - and therefore heritable - psychological modules (Pinker, Mind *passim;* Tooby and Cosmides, 34). Again, the evidence for this claim is too vast to discuss, and I can only point readers to the review of the current state of research in How the Mind Works. Consider, as an point of entry to this field, the function of the human eye and its connection to the brain. Everyone from William Paley to Richard Dawkins has hailed this organ as a spectacular example of precise and complex engineering too intricate to ever have arisen by mere chance. The eye automatically focuses and adjust to changing light conditions, lubricates itself, protects itself from trauma, self-repairs, and generates visual information in full color and of an incredibly high quality. But, as Tooby and Cosmides point out, there is a delicious irony here in supposing that the highly specialized physiology of the eye exists only to deliver a visual signal to a non-specialized general purpose information processor. To the contrary, we have every reason to think that the upstream processing of visual input in the brain must be every bit as sophisticated and content-specific as the organ whence the input originated (Adapted Mind, 55-9). And Stephen Pinker points out that many problems in vision are formally insoluble unless the mind employs context-specific heuristics to resolve otherwise ambiguous visual

phenomena. Physical movement, use of language, ability to discern relevant from irrelevant data, ability to comprehend social interactions, to name but a few, are likewise all exceedingly complex tasks which humans do effortlessly, even though four decades of research into artificial intelligence have failed to emulate human performance in these areas. Moreover, the fact that physical damage to specific sections of the brain results in impairment to specific human skills (Damasio) suggests very strongly that the brain is not a general purpose computer which acquires its abilities by simple learning them, but a complex whole composed of numerous interrelated problem-specific modules.[84]

Modularity is an increasingly important part of evolutionary psychology, but its origins lie in the work of Noam Chomsky and Jerry Fodor, both of whom however avoided an evolutionary view of modules. Modularity, in the hands of evolutionary psychologists, is an attempt to explain how evolutionary processes could shape the human mind, a heuristic in making predictions about how it might operate in a given domain, and a corrective the behaviorist approach of sociobiologists, who offer adaptive explanations only for behaviors (Carruthers and Chamberlain Human Mind, 3-4). Although theorists have not reached widespread understanding on what exactly modules are, there are a few features common to most accounts: modules are domain-specific, informationally encapsulated, generally inaccessible to other processes, are adaptations, and comprise a good deal of the neural architecture (Murphy and Stich, "Darwin in the Madhouse", 64).

There are compelling theoretical reasons to accept such a view of the mind. Stephen Pinker points out that the minimal criteria necessary to impute rational

[84] This should not be confused with the claim that different areas are "responsible" for different functions. Typically, any given mental function will involve many sectors of the brain, but the loss of a given area may make that function impossible to realize.

behavior to an agent is that the agent apparently has the capacity to follow rules (or principles) which connect an agent's action with reality, and that the agent uses these rules to pursue some goal, and to be able to do so in different ways when obstacles impede her (Mind, 61). Rationality thus connects belief and desire through a large set of rules which are adaptive and flexible. Brains have the power to represent information (about beliefs and rules and desires) with symbols, and these symbols, in the right context, have causal powers which can create more information, and finally action (Mind, 65-7).

According to Pinker, the mind represents information through the use of modules - task-specific agents that process one sort of information and which are themselves composed of smaller, simpler, modules, and so on, until we find at the bottom of the computational heap, the simplest of logic switches. It is each module's task to process information appropriately and then to pass it on to the next module. And such a modular theory of the mind is compatible with the way in which computer programs are written: repeated tasks are accomplished by modularized subroutines which "hide" simpler, much-used processes, and these routines can be reinvoked at will without having to rewrite the code. The human body is also organized on similar lines: each organ of the body performs a specialized function, and its parts are further specialized to fulfill their own contributing purposes, and so on, unto the level of the cell, where even there, specialized structures perform specific functions (Mind, 90-92). But mental modules do not all process information in the same way. The mind is apparently capable of widely differing forms of representation, and of processing the same information using different representations depending on the context. For example, if people are asked to identify letters of the alphabet as being the same, they will interpret the series "AA" as a series of images and compare them for

visual similarity. But when presented with "Aa", they will interpret each letter as "the letter A" and infer their similarity in a different manner (and will do so more slowly). But when there is a sufficient time period between one "A" and the next, people convert the earlier visual image to an alphabetic representation, and the slight difference in speed disappears even between identical uppercase "As" (Mind, 89). And Pinker's work on the use of regular and irregular verbs shows that humans cannot be using just one sort of representation to conjugate verbs (Words and Rules, *passim*).[85]

But how exactly might the mind represent the rules of grammar or the rules of rationality? Nozick's suggestion is that while philosophers strive to derive compact sets of principles to guide reason, this need not be the way the human mind works

> If the rationality of a belief, however , is a function of the effectiveness of the process that produces and maintains it, then there is no guarantee that optimal processes will employ any rules that are appealing on their face. Those processes instead might employ scorekeeping competition among rival rules and procedures whose strengths are determined (according to specific scoring procedures) by each rule's past history of participation in successful predictions and inferences. None of these rules or procedures need look reasonable on their face, but, constantly modified by feedback, they interact in tandem to produce results that meet the desired external

[85] Some examples of proposed modules are: Chomsky's language acquisition module, the theory of mind module that allows humans to attribute mental states to others, and which autistic persons apparently do not have (Pinker Mind, 329-33), Stephen Pinker's various modules which conjugate regular and irregular verbs (Words and Rules), John Tooby's and Leda Cosmides' social contract cheater detection module ("Cognitive Adaptations for Social Exchange"), Dan Sperber's symbolic interpreting mechanism (Symbols), etc. These comments are only intended as brief overview - and not a complete defense - of this theory. Carruthers' and Chamberlain's recent Evolution and the Human Mind is recommended for detailed exposition and defense of the modular theory of mind.

> criteria (such as truth). The theory of that process, then, would not be a small set of rules whose apparent reasonableness can be detected so that a person then could feasibly apply them but [sic] a computer program to simulate that very process (Rationality, 75-6).

If our belief-making processes did in fact work in this way, then, says Nozick, philosophy's search for tightly consistent sets of action- and belief-guiding principles will quickly be rendered obsolete by cognitive scientists, artificial intelligence specialists and others who will build computers capable of interpreting data in a way more akin to the way the mind does and thereby able to represent answers which are more amenable to human understanding - but those computers will not be following any "rules" which humans could easily understand or express (76-7).

This observation marks a striking departure from what has been dubbed the GOFAI (for "Good Old-Fashioned Artificial Intelligence") school of the computational model of the mind. On this view, the mind performs operations by manipulating logical rules serially, just as the garden variety von Neumann computer does. In contrast, a connectionist model suggests that the mind does not process information by the explicit syntactical manipulation of symbolic data. Rather, mental representations are derived form the total state of the computing apparatus, and the information is distributed throughout the system, rather than being stored in a discrete location as a symbolic output (Ramsey, "Connectionism, 186). The central elements of these connectionist nets are:

> (a) simple processing units or nodes, which sum the incoming activation, following a specified equation, and then send the resulting activation to the nodes to which they connected,

(b) equations that determine the activation of each node at each point in time, based on the activation from other nodes, previous activation, and the decay rate,
(c) weighted connections betwe. . the nodes, where the weights affect how activation is spread, and
(d) a learning rule that specifies how the weights change in response to experience (Read and Miller, ix).

The simplest forms of connectionist nets are "feed-forward" ones in which activation levels are set once, and data flows only from input nodes to output nodes, and never the other way . The other chief type of connectionist net is the "feedback," "interactive," or "back-propagation" nets in which outputs are compared to a desired result, and connection weightings are adjusted slightly until outputs are brought into equilibrium (Read and Miller, ix-x).

As an illustration, consider the connectionist net "NETtalk" which consists of 203 input units, 26 output units, and 80 "hidden" units interposed between the input and output arrays. Each input unit is connected to each hidden unit, and each hidden unit to each output unit, for a total of 18, 629 connections. NETtalk can mimic spoken language by comparing speech inputs and and processing that signal to create an output, and then modifying that output until it matches the input. NETtalk displays many of the features which defenders say shows the superiority of connectionist schemes as models of the human mind: NETtalk can learn new dialects from scratch with no initial programming, its performance improves with experience, it employs representations of hierarchically ordered phonetic rules similar to those identified in phonetics, and disabling particular elements ("lesioning") results in localized disabilities qualitatively similar to those found in humans (Boden, 15-16). Connectionists also point out that

connectionist nets are more "biological" than von Neumann architectures, since their multiple connections emulate to some degree the connections between actual neurons, are especially adept at pattern recognition, degrade gracefully as human performance does when the system is damaged slightly (Read and Miller, x), do not rely on a "top-down" design in which each computational function is over-efficiently specified for each element, do not rely on explicit rules (Dennett "Mother Nature", 22-25), and can function even when input contains a considerable amount of "noise."

Nonetheless, critics charges that connectionism suffers many shortcomings in that it cannot easily emulate many human processes, and researchers remain divided as to whether connectionism most closely aligns itself with nativist (the mind contains much innate programming) or empiricist (little or no innate programming) views of the mind, whether connectionist nets in the mind are localized or distributed, and whether connectionism supports an eliminativist view of the mind or not (Ramsey, 186-7; Dennett "Mother Nature," 22-3, 27-8; Read and Miller, x-xi). I think an extended discussion of all these issues would take us rather far afield, and my major purpose here is simply to propose that a connectionist model can stand as a plausible model of how mental modules might function.

These four considerations (twin studies, the existence of human universals, the complex functionality of psychological traits, the necessity for and experimental evidence of domain-specific modules) taken together, provide considerable support for the claim that many psychological traits, especially those closely responsible for generating behavior that has historically affected fitness, are indeed heritable. It follows from all this that since many aspects of human psychology (including sensory capacities, desires and dispositions, emotions,

and reasoning) are heritable, variable, and contribute differentially to fitness, those psychological features which would have increased (or optimized or maximized) fitness in the EEA would therefore have been preserved by natural selection. And those psychological features which have thus endured to the present day can be reverse-engineered to determine their adaptive value in the EEA.

Some Implications of Evolutionary Psychology

Social constructionists hold that because humans have only "unspecialized and undirected" drives, this means that "the human organism is capable of applying its constitutionally given equipment to a very wide and, in addition, constantly variable range of activities (Berger and Luckmann, 48)." Evolutionary psychologists argue that this is exactly wrong and that is there is no reason to think that task-specific dispositions which enhanced the fitness of our distant ancestors should have disappeared from the human psyche . Evolutionary psychology holds, human minds are extremely well-endowed with these innate mechanisms, they are for that reason massively adaptable, able to manipulate their environment, and able to create highly specialized and complexly arranged cultural artifacts which serve various adaptive ends. Again, a computer analogy is appropriate here. The number of possible different outputs of which a computer is capable is limited by the number of possible computational states the program is capable of generating. These in turn are limited in part by the number of different sorts of inputs the program can recognize. (Think of a very simple word-processing program which is incapable of recognizing commands to change typeface styles, for example.) Obviously a large and functionally complex program is capable of more (and more interesting) responses than a

smaller and simpler one.[86] Likewise, humans possess numerous psychological mechanisms that work in conjunction with each other and this is why we are capable of such a wide variety of behaviors, compared to ants, who (presumably) have less (not more!) innate "programming", and correspondingly more "robotic" behavior. On this account, the diversity of human behavior - compared to ants or dogs, say - is evidence of more programming, not less.[87]

> Any time the mind generates any behavior at all, it does so by virtue of specific generative programs in the head, in conjunction with the environmental inputs with which they are presented. Evolved structure does not constrain; it creates or enables (Tooby and Cosmides 39).

Like their ancestors, humans do not seek directly to maximize or even to increase fitness. Instead, they strive to achieve numerous sub-goals (i.e., those necessary for survival and reproduction: eating, avoiding danger, mating, child rearing, successful social integration, etc.) which would have increased fitness in the EEA. Culture is thus - as Richard Dawkins puts it - part of the human gene's "extended phenotype," the product of an evolved and specialized psychology operating in a given environment (Extended Phenotype). Thus culture is not an entity *sui generis* that creates and perpetuates itself untouched by individual human psychology, as some social constructionists have claimed. Moreover, since not all parts of a human's environment (including its culture) will constitute its developmental environment, and the human developmental plan

[86] This example is meant to be illustrative, and certainly not definitive, of the way the human mind works. The reader should resist the erroneous conclusion that the mind is a rigidly deterministic input-output machine.

[87] This analogy should be read to mean that I am likening the human mind to a computer that possesses large computational powers. rather, the analogue is a computer that has numerous, domain-specific subroutines a that allow it to respond to a variety of inputs in specific ways. Of course, this degree of specialized programming may imply large computational powers as well.

directs the mind to construct itself by interacting only with some (relevant and regular) features of the environment, what counts as an environmental influence is itself, to a large degree, shaped by the way in which genes direct humans to seek out particular aspects of their environment.[88]

Objections to Evolutionary Psychology

Such a short account cannot do full justice to the complexity of the debate around evolutionary psychology. Still, I want to consider a few objections to EP which are common in the literature. These objections fall naturally into two camps - those objections which count particularly against psychological (as opposed to physiological) applications of evolutionary theory and those that, if valid, would count against evolutionary theory broadly construed (even though their proponents may not intend them to be so construed). This latter class, I think, is far less worrisome than the objections which are directed at evolutionary psychology itself, and accordingly, I will devote much less space to them. In each case, I offer an uncontroversial *reductio* which uses an example drawn from evolutionary accounts of physiological traits.

Some Cases of Special Pleading

For example, some people believe that not one or a few, but very many genes, interacting in a complex and unpredictable ways with each other and their environment, are responsible for the equally complex and unpredictable processes which create the brain. This process is so complex and so inextricably

[88] See Tooby and Cosmides for extended discussion.

subject to random and multifarious environmental contingencies that it is impossible to say that the genome can have any determinate effect (in the sense of imposing any definitive nature) on the brain. Accordingly, interactionists reject both biological and environmental explanations of biological function:

> A second, more pluralistic, response to biological determinism is interactionism. According to this view it is neither the genes nor the environment that determines an organism but a unique interaction between them. Interactionism is the beginning of wisdom .... There are no generalities that hold consistently about the ways in which different genotypes will develop differently in different environments. It all depends. (Rose et al, 268)

Stephen Pinker points out that interactionists, insofar as they are unwilling to give a detailed account of any of the specific mechanisms that play a role in biological function, cannot give adequate explanations of those functions at all, and this is evident if we attempt to formulate interactionist explanations for other complex systems:

> The behavior of a computer comes from a complex interaction between the processor and the input.
>
> When trying to understand how a car works, one cannot neglect the engine or the driver. All are important factors.
>
> The sound coming out of a CD player represents the inextricably intertwined mixture of two crucial variables: the structure of the machine, and the disk you put into it. Neither can be ignored. (Mind, 32.)

As Pinker observes, "[t]hese statements are true but useless - so blankly uncomprehending, so defiantly uncurious, that it is almost as bad to assert them as to deny them (Mind, 32)."[**] Once we recognize that interactionists must be willing, at least in principle, to specify the actual mechanisms that allow an organism to interact with its environment, we can then see that human rationality is not simply a general-purpose capacity for abstract thought which we press into service to solve particular problems, but is rather a set of specialized tools for solving specific problems which we have repeatedly encountered in the past. And this realization will have profound implications for any philosophical understanding of rationality.

"Complexifying" and interactionist accounts are not only explanatorily sterile, they seem at odds with the ways in which biological processes create physical features. No matter how complex these processes are, no matter how dependent they are upon on unpredictable and variable extra-genetic effects, no matter how incomprehensible and unpredictable they may appear from the standpoint of human understanding, these processes, no matter what their nature may be, for the most part reliably construct human hearts, stomachs, muscles, etc., which for the most part reliably perform the same functions in the same way no matter which human body or which environment they happen to be located. Since the

[**] This is no overstatement. Consider that Richard Lewontin (who, with Steven Rose, co-authored the passage above) is perhaps the most able and articulate defender of interactionism and the director of a prestigious Harvard research lab. But Michael Ruse believes that Lewontin's early scientific accomplishments were effectively ended when he rejected reductionism (on political grounds) and adopted "dialecticism" in its place - a research heuristic that proved remarkably sterile. "As an active scientist," writes Ruse, "Lewontin has produced virtually nothing.... His own greatest scientific achievements were the epitome of the reductionistic approach ... he has embraced a philosophy that condemns his science (Globe and Mail)." Indeed, Lewontin has published only four scholarly publications within his discipline in the quarter century since 1974's The Genetic Basis of Evolutionary Change (Lewontin "Lewontin").

complexity/contingency argument is obviously inconsistent with the reliable and highly detailed similarity (between individuals) of function of these organs, I am at a loss to understand why it is thought decisive where brain functions are similar. Does this mean that critics of evolutionary psychology deny that rationality is an adaptation? Lewontin argues that the human genome is simply too small to carry sufficient information to code for the construction of individual neuronal connections. "Once we admit that only the the most general outlines of social behavior could be genetically coded, then we must allow immense flexibility depending on particular social circumstances (Biology 72)." Evolutionary psychologists agree that of course we do have considerable flexibility, but suggest that this points to more, rather than less, coding. And Lewontin, of course, never specifies exactly which social responses are coded. And yet the human genome - as puny as Lewontin presents it - undoubtedly codes for some very complex procedures - vision, cell building and reproduction, blood clotting, the antibody system, to name but a very few - are all so complex that we barely understand how many of them work in any detail. Why then should we accept that it is too small to code for complex social mental functions?

Another objection runs as follows: if rationality was in fact an adaptation, then carrots (and presumably all other organisms) would be rational. But they are not. Therefore rationality is not an adaptation.[90] The obvious suppressed premise here is that if a trait is an adaptation, it must be expressed universally. If we understand this premise to apply to all biological traits, it is patently absurd: nothing in evolutionary theory compels us to assert that because wings, opposable thumbs, and photosynthesis are non-universal, they cannot, for that very reason, be adaptations. If we understand the suppressed premise to apply

[90] Several subscribers to the e-mail mailing list PHILOSOP made,and vigorously defended, this point, 2000.

only to mental traits, it implies a dualism which must be recognized and defended. Either way, the idea that adaptations must be universal is itself implausible. This interpretation ignores the fact that the possibility of acquiring a certain adaptation depends on the organism's needs (which in turn are determined in part by the adaptive niche in which it lives) and on the possibility that it can acquire such a trait by numerous, small, successive adaptive steps, each of which provides an increase in fitness. That is, some adaptations might confer a benefit on an organism, but the organism could not acquire them because the intervening stages of adaptation between not having the adaptation and having it confer lower fitness on the organism.[91] Barring huge mutations (the prospect envisioned by saltationism) or such a temporarily unhelpful mutation piggybacking on some much more beneficial adaptation, organisms generally cannot suffer a temporary loss of fitness to acquire a greater future fitness. So an organism cannot acquire a beneficial trait unless it it is evolutionary accessible to it. Some adaptations (those controlling reproduction, for example) might pass this test and be expressed universally. But there is no reason to think that all traits (including rationality) must pass this test.

Again, Elliot Sober points out that knitting correlates strongly with having two X chromosomes, but that it would be absurd to infer from this fact that double X chromosomes contain genes "for" knitting (Biology 186-8).[92] The BSSRS Sociobiology Group similarly point out that avoiding the wrath of the Inquisition surely enhances fitness for mediaeval European peasants, but this is no evidence that the disposition to do so is an adaptation (Birke and Silvertown 120, 122). But

[91] For example, it might benefit a cat to grow a mousetrap on its paws. But the mousetrap will likely be useless until all its components are in place and functional. In the meantime, cats bearing only a few useless parts of the mousetrap will have less fitness than their forebears.

[92] This is a poor example to make Sober's point, since knitting is apparently more common practiced by men in some cultures.

exactly similar arguments can be made about traits which are not the product of conscious decisions. Suppose, for example, that having at least one Y chromosome correlates strongly with wearing large running shoes. Obviously, Y chromosomes do not "code" for wearing large running shoes, but this does not preclude our suggesting that Y chromosomes don't code for some more general feature (i.e., greater body mass) which entails having larger feet, and hence a need for larger running shoes. Similarly, the evolutionary psychologist need not admit that double X chromosomes code for knitting, but this fact does not prevent her from suggesting some other broader psychological generalization which might explain different knitting preferences. A parallel answer can also be made to the BSSRS Sociobiology Group. Consider that being able to digest hamburgers might enhance fitness in a given environment, but the evolutionist can just say that a disposition to eat hamburgers is not due to any adaptation specifically for hamburgers, but for a general group of foods, to which hamburgers are closely enough related to be digested. So there is an adaptive explanation for hamburger digestion, but it involves positing a more general digestive ability, and does not commit the evolutionist to positing a "hamburger-specific" adaptation. In the same way, the evolutionary psychologist can suggest that those who avoided the Inquisition were perhaps motivated by an adaptive mechanism, but that the adaptation might be a more generalized disposition, say, to avoid being harmed by powerful conspecifics, and not an adaptation selected specifically to avoid the Inquisition. Notice that exactly similar problems arise in the case of non-human behavior: when a frog reacts to a fly in its environment by catching it with its tongue, is its disposition to catch only members of the genus *Drosophila*, to catch flies, to catch flying insects, or simply to catch flying black blobs? Even if the precise disposition cannot (even in principle) be specified, this in no way counts against the claim that such a

disposition might be an adaptation. If this problem does not render adaptationist explanations vacuous, why would anyone think it renders EP adaptationist stories vacuous? In fact, an exactly similar (and equally irrelevant) problem can be posed for purely social explanations for behavior: social determinists can argue that the alienating forces of North American culture might create a general disposition for male violence, but they are not thereby committed to saying (for example) that society "programmed" Charles Manson to kill the specific individual named Sharon Tate.

One final example should make it clear that many objections to EP are really objections to evolutionary theory itself. It is no secret that the putative social implications of EP are those which attract the most attention and the strongest objections. For example, the BSSRS Sociobiology Group contends that "... dominance is a social relation between individuals rather than a property they possess" and therefore, they conclude, not a heritable property (Birke and Silvertown 140, 150-51). But redefining "properties" as "relations" (and therefore non-heritable) leads to absurd consequences. After all, camouflage and digestion are also "relations" between an organism and its environment, but they are nonetheless plainly heritable. I see no reason to think that the BSSRS group is entitled to any special pleading in the case of psychological traits. Again, the problem may simply be one of properly identifying the trait under discussion in an explanatorily useful way.

So none of these objections offers any substantive reason to doubt evolutionary psychology. What is striking about their use is that so few people recognize that they count with equal force (that is, none at all) against evolutionary explanations of non-psychological traits. These objections are, in fact, strikingly akin to the sorts of arguments which creationists offer against evolutionary

theory. This perhaps explains why critics of evolutionary psychology are sometimes unflatteringly described as "the new creationists."

I turn now to two more substantive objections against evolutionary psychology. First, evolutionary psychology is deterministic. Second, it carries deeply disturbing political consequences that speak against it in various ways.

The Bogeyman of Biological Determinism

The most common response to any suggestion that genes or evolution might influence any aspect of human behavior or psychology is to label it as "biological determinism." The first thing to notice about this designation is that it is a term of abuse. Critics (Rose et al, Birke and Silvertown, Lewontin Ideology, Tuana, Purdy, Rodd, etc.) describe anyone who posits biological explanations for human behavior as a "biological determinist". In fact, Richard Lewontin holds that "Except for a brief interruption around the time of the Second World War, when the claims of Nazism made claims of innate inferiority extremely unpopular, biological determinism has been the mainstream commitment of biologists (Ideology 26)." But, in my (admittedly limited) readings, I have yet to find anyone who describes him/herself as a biological determinist. This, surely, is a clue that critics of evolutionary psychology fundamentally misunderstand how practitioners of evolutionary psychology understand the project. For example, some critics unhesitatingly class Richard Dawkins as one of the worst of the biological determinists (Rose et al, 8; Dusek). This is somewhat inexplicable since Dawkins opens The Extended Phenotype (which he counts as his most important work) by forcefully denouncing biological determinism as a myth - a frightening

and oft-repeated myth, to be sure - but one no more to be believed than World War II rumors that Russians had invaded Scotland (9).[93]

Perhaps some evolutionary psychologists are (closet) determinists. Which is to say: they believe that nature is destiny and that any genetic influence on one's psychology must inerrantly express itself phenotypically paying no heed to external or internal forces. And they are forced by this into a resigned acceptance of the imperfection and non-perfectibility of humans and of the non-progressive politics that such a view of humanity must entail.

I think, however, that the majority opinion does not hold this view, and that evolutionary claims about human behavior need not entail biological determinism - whatever the phrase might mean. We need not choose between radical genetic determinism and radical environmental determinism. More plausibly, those aspects of human behavior which are affected in specific ways by genetic influences will express themselves statistically (and not in an absolutist, essentialist manner) within a population, and will present themselves as conditional responses to specific environmental stimuli, where those responses themselves are subject to further modification by other evoked responses.

But what exactly is biological determinism? According to one widely-regarded source, biological determinism is the doctrine that

> ... human lives and actions are inevitable consequences of the biochemical properties of the cells that make up the individual; and these characteristics are in turn uniquely determined by the constituents of the

[93] Val Dusek, who perhaps fails to interpret this example as an analogy, denounces it as an "unrelated" story, and accuses Dawkins of gratuitous red-baiting.

> genes possessed by each individual. Ultimately, all human behavior - hence all human society - is governed by a chain of determinants that runs from the gene to the individual to the sum of the behaviors of all individuals.(Rose et al 6)[94]

We might describe this as hard biological determinism, since it admits of no role for environmental influence whatsoever. And it is also a form of global biological determinism, since it suggests that it is the full panoply of biological genetic forces which, acting in concert, determines an individual's behavior, rather than a single gene controlling a single behavior. Put this way, biological determinism is a straw man, and on this there seems to be a broad consensus, even among those traditionally counted as biological determinists:

E. O. Wilson:

> Each person is molded by an interaction of his environment, especially his cultural environment, with the genes that affect social behavior. Although the hundreds of the world's cultures seem enormously variable to those of us who stand in their midst, all versions of human social behavior together form only a tiny fraction of the realized organizations on this planet and a still smaller fraction of those that can be readily imagined with the aid of sociobiological theory. (Wilson, Human Nature 18-19)

Daniel Dennett:

---

[94] 94 Strictly speaking, what Rose et al are defining is not biological determinism, but genetic determinism, since there are other biological forces besides genes that contribute to "biological determinism," if such a force exists. (Sober 1993, 192). There is often further confusion between "innate" "inherited" and "genetic", none of which are quite the same thing. I am here only concerned with genetic determinism.

> Whereas animals are rigidly controlled by their biology, human behavior is *largely* controlled by culture, a *largely* autonomous system of symbols and values, growing from a biological base, but growing indefinitely away from it. (Idea, 491)

Martin Daly and Margo Wilson:

> [I]t is a widespread misapprehension that biological approaches to the study of violence are narrowly 'deterministic" in some way that is antithetical to the analysis of social and circumstantial influences.(Daly and Wilson, Homicide 296)

Stephen Pinker:

> [N]atural selection is not a puppetmaster that pulls the strings of behavior. It acts by designing the generator of behavior: the package of information-processing and goal pursuing mechanisms called the mind. Our minds are designed to generate behavior that would have been adaptive on average in our ancestral environment, but any particular deed done today is the effect of dozens of causes.(42)

Michael Ruse:

> We are not ants. Much that we do socially requires learning, and ... we seem to have a dimension of freedom, of flexibility, not possessed by the ants - which is just as well, biologically speaking. Genetic hardwiring is just fine and dandy, as long as nothing goes wrong. But when there are new challenges, it is powerless to pull back and reconsider.(Naturalism

239)

R. D. Alexander:

> To say that we are evolved to serve the interests of our genes in no way suggests that we are obliged to serve them. (cited in Bradie 203)

Richard Dawkins:

> [Biological determinism] is pernicious rubbish on an almost astrological scale. Genetic causes and environmental causes are in principle not different from each other. Some influences of both types may be hard to reverse; others may be easy to reverse ... What did genes do to deserve their sinister, juggernaut-like reputation? Why are genes thought to be much more fixed and inescapable in their effects than [environmental influences such as] television, nuns, or books? (Extended Phenotype 13)

John Tooby and Leda Cosmides:

> The fact that humans in ordinary environments reliably develop a clearly recognizable species-typical architecture should in no way be taken to imply that any developed featured of any human is immutable or impervious to modification or elimination by sufficiently ingenious

> ontogenetic intervention... In contrast, Standard Social Science Model[95] advocates, such as [Stephen Jay] Gould, tend to equate evolved biological design with immutability without any logical or empirical warrant. As Gould expresses his rather magical belief, "If we are programmed to be what we are, then these traits are ineluctable. We may, at best, channel them, but we cannot change them by will, education, or culture." ("Foundations" 80)

It is hardly overkill to cite these numerous sources. As recently as 1995, the well-known Harvard biologist Richard Lewontin told Canadian audiences through a series of CBC Massey lectures and a subsequent book based on those lectures that "It is a fallacy of biological determinism to say that if differences are in the genes, no change can occur (Lewontin, Ideology 30)." Lewontin offers not a single example of any reputable biologist who holds this belief.[96]

In short, and as evolutionary psychologists endlessly repeat, genotypes do not determine phenotypes. Hard global biological determinism is wrong because it does not recognize the role of the environment in mediating genetic effects, and this is why (so far as I can see) it has no reputable adherents. Michael Ruse, one of Canada's most eminent philosophers of science, says, "I know of no one who

---

[95] "The Standard Social Science Model" is Tooby's and Cosmides' designation for what they take to be the cluster of ideas that has characterized the dominant social science model of human development and behavior, and the main alternative to evolutionary psychology. The dominant ideas of the SSSM can be characterized as follows: human infants are all very much alike but adults differ greatly between cultures. Therefore culture is the main or only determinant of within-group similarities and between-group differences. Culture is transmitted by learning, a unitary process that is well-understood and requires no special complement of innate structures. Culture itself is an emergent property not dependent on any specific features of human psychology. In fact, humans have few or no instincts, and if they did, those instincts could only produce robotic behavior (1995, 31-2). Tooby and Cosmides' 1995 "The Psychological Foundations of Culture" is perhaps the best account of the differences between evolutionary psychology and the SSSM.

[96] Nonetheless, Lewontin's work is taken seriously enough that his book has even been used in philosophy courses.

is consistently a hardline [biological] determinist, metaphysical or methodological... (Paradigm, 164)." In this sense, the gap between biological, social, and dialectic explanations of human behavior and psychology has shrunk somewhat. In the words of one observer, "We're all interactionists now."[97]

Having said all that, I must now take some of it back. In the case of some genes, their effect is such that no environmental change will affect their expression. Huntingdon's chorea, for example, is one such effect. Huntingdon's inflicts increasing chorea, atheosis, mental deterioration, and finally death, on its victims (Berkow 313-4), all of which result solely because of a minor mutation which repeats the sequence "CAG" on chromosome 4:

> The cause is in the genes and nowhere else. Either you have the Huntingdon's mutation and will get the disease or not. This is determinism, predestination, and fate on a scale of which Calvin never dreamed. It seems at first sight to be the ultimate proof that the genes are in charge and that there is nothing we can do about it. It does not matter if you smoke, or take vitamin pills, if you work out or become a couch potato. The age at which the madness will appear depends strictly and implacably on the number of repetitions of the "word" CAG in one place in one gene. If you have thirty-nine, you have a ninety per cent probability of dementia by the age of seventy-five and will one average get the first symptoms at sixty-six; if forty, on average you will succumb at fifty-nine; if forty-one, at fifty-four; if forty-two, at thirty-seven; and so on until those who have fifty repetitions of the "word" will lose their minds at roughly twenty-seven years of age. The scale is this: if your chromosomes were long enough to stretch around the equator, the difference between health

[97] J. S. Roberts, p.c., Learneds, May 2000.

> and insanity would be less than one extra inch .... Huntingdon's disease .... is pure fatalism, undiluted by environmental variability (Ridley 56, 64)

Ridley says this example only "seems" to warrant belief in genetic determinism because, of course, many genes do not work this way. Genetically-shaped dispositions which control behavior, for example, are more likely to be conditional ("if thirsty, drink") rather than categorical and their influence on an individual at any given time will depend on environmental influences and their relative importance compared to the perceived importance of other dispositions. And, moreover, because traits which have been favored by natural selection must be variable between individuals, there is no reason to think that all individuals will express a trait to the same degree, even in the same environment. Tanning, for example, is an adaptive mechanism, but no-one thinks that all individuals (even those who with close genetic ties) will tan in the same way. And it would be the height of folly to suggest that, because we found one person who could not tan, therefore there is no widespread human adaptation which expresses itself phenotypically as tanning.

This is, however, exactly the strategy that Rose and Lewontin employ repeatedly to defuse any claims for widespread genetic dispositions to a certain behavior. They assume in every case that if there is such a gene (say, a gene that creates a disposition to xenophobia), it must express itself invariantly and universally. All then that is needed to refute such claims is the existence of a single phenotypic counterexample (i.e., someone who is not xenophobic) and this is always easy to find. If an exception can be found, then they assume there is no genetic effect.[98] Rose and Lewontin evince a similar misunderstanding of claims for between-group differences, such as those which are claimed to exist between men and

[98] John Tooby and Leda Cosmides argue that Margaret Mead and Emil Durkheim employ a similar strategy (Adapted Mind, 43).

women. They interpret a claim for some sex-related difference X (say, that women express X more than men do) to mean that every woman will express X more than every man will. Accordingly, it is ridiculously easy for Rose and Lewontin to find some individual that does not conform to this putative sexual stereotype (Rose et al 138, 146, 246; Lewontin, Ideology 65-66, 67-80). These are such needless, uncharitable, and simplistic misreadings, that they are barely worth refutation.

So if "hard" biological determinism is a bogeyman which no-one either takes seriously or should take seriously, it then seems reasonable to ask if some "softer" version of the theory is any more plausible. That is, if we interpret claims for genetic dispositions to a behavior of one sort or another as claims that can be expressed as conditional universals or statistical correlations, how then may we interpret these claims?

Two thorny theoretical problems arise here. The first lies in way we parse claims about genetic influence. Suppose we say that 50% of the difference between two groups of individuals' disposition to a certain cancer is due to a sheared genetic influence. It is tempting to interpret this claim as saying that a member of one of those groups inherits 50% of his risk of cancer from his parents. But this is a misleading way of looking at the problem. We all understand that a child's height is influenced in part by the genes she inherits from her parents and in part by her environment. But it would be nonsensical to say that she inherits 93 cm of her height from her parents and 46 cm from her environment. In precisely the same way, there is no genetic component to one's personality (or susceptibility to cancer) which is distinct from an environmental component. Rather, we can only suggest that a genetic (or environmental, for that matter) difference between two

otherwise similar populations is responsible for some degree of difference between these two populations.

But even this account may be too simplistic. Richard Lewontin points out that genetic effects may vary between different environments (29). To illustrate this, imagine that two species of wheat, A and B grow in the same environment and that A produces 20% more wheat than B. It seems here that this 20% difference is entirely due to their differing genetic makeup. But this need not be so. If A and B are grown in a second environment (say, where they are heavily fertilized) it is entirely possible that B will outproduce A by 30% because B responds more to fertilizer than does A. So, on this account, we cannot even quantify, in any absolute sense, the difference in yield which is due to genetic differences between A and B, because this difference itself depends on environmental effects.

The logical extension of this to evolutionary psychology is that one cannot specify categorically what phenotypic effects any supposed genetic disposition will have, because these effects are also dependent on environmental variables. And it is obvious that humans are much more able to manipulate their environments and to thereby change their expressed personalities in many more ways than, say, wheat plants can. Lewontin reports for example that some people claim that as much as 80% of differences in I. Q. between individual children is due to genetic differences and that therefore this difference cannot be eliminated. But, Lewontin claims, this is "completely fallacious" apparently because all of this difference can be eliminated by environmental measures such as providing children with electronic calculators. Similarly, putative differences in "strength and physique" between men and women will "disappear from practical view" when humans are provided with power lifts, power steering, etc. (Lewontin, Ideology 29-30). There is surely room to wonder just what Lewontin

can mean by "intelligence" and "strength" in this context and to ask if his claim that "natural" measures of these properties are illusory is valid. But his greater point can stand: genetic effects do seem to vary between environments, and this must be relevant to our study of human psychology.

However, we should remember that we have already seen in one case - a limit case, to be sure - that this is simply false. Huntingdon's chorea kills all individuals who survive to a certain age, and spares all others and this is an effect which is not affected by environmental considerations at all. Against this, J. S. Roberts points out that Huntingdon's does not kill those who die (or kill themselves, as many Huntingdon's gene-bearers do) before the onset of the disease, and that this "important" exception powerfully supports the claims that genetic and environmental influences cannot be separated (p.c., Learneds, 2000). To respond in this way, it seems to me, to is grasp at a theoretical straw. Consider this analogy: in most environments, a Macintosh computer running Operating System 9.0.4 will behave in ways markedly different from an otherwise identical Mac running Operating System X. But this difference, I concede, will undoubtedly disappear if a nuclear explosion vaporizes both computers. Nonetheless, it would be pointlessly dogmatic to insist that one cannot therefore distinguish between the effects of the software and the effects of the environment when comparing these two computers. The moral here is that while absolute claims for a given genetic influence are perhaps illegitimate, this in no way prevents us from predicting its influence across a range of relevantly similar set of environments. And plainly, when a given genetic effect is observed in many or all observable human environments, this suggests that this genetic influence is not strongly affected by most environmental changes at all.

The second theoretical problem I wish to consider arises from the fact that evolutionary explanations of psychological traits are adaptationist stories. Sober defines adaptationism as the belief that "most phenotypic traits in most populations can be explained by a model in which selection is described and nonselective processes are ignored (Biology 122)." Of course, other evolutionary forces (founder effect, genetic drift, environmental noise[99], etc.) do play roles in evolution, but this does not count against the point that wherever we encounter complex functional traits (binocular vision, flight, blood circulation, etc.), the best explanation is overwhelmingly likely to be an adaptive story. Adaptationism is therefore the default heuristic in seeking explanations for these traits. The trouble with adaptationism, however, is simply that it generates too many hypotheses. For almost any human physical trait - bipedalism, hairlessness, large brains - numerous adaptive stories can be invented which purport to show why the trait is the optimal result of some selective force or another. All too often, however, evolutionists are unable to discover evidence that supports some adaptive "just-so" story, or even to imagine a way in which an adaptive hypothesis could be tested. Obviously the same is true when adaptive explanations are offered for human behavior. Moreover, when humans fail to exhibit some behavior which is claims to be caused by some putative adaptive psychological disposition, one can always posit some other *ad hoc* mental faculty which has overridden the first, and this reduces even further the plausibility of the adaptive explanation.

Elliot Sober's suggestion is that if optimality models are too easy to construct, biologists should make them harder. He offers this example, drawn from G.

[99] "Founder effect" describes the consequences due to the founding of a new group with very few initial members. If one member of this small group has a rare gene, it will automatically be far more prevalent in that group than it may have been in a large population. "Genetic drift" marks the random change in genetic frequencies through time, especially in small groups. Definitions from The Cambridge Encyclopedia of Human Evolution, 462, 463. "Environmental noise" denotes the genetic effects of random environmental events (earthquakes, forest fires, etc.) to which natural selection cannot respond.

Parker's research on the mating behavior of dung flies. Male dung flies fertilize more eggs the longer they copulate with females, but the number of eggs fertilized is not directly proportional to the length of time copulating. In fact, males receive diminishing returns the longer they copulate. Moreover, longer copulations times reduce the amount of time males have to seek out other females. As it happens, males copulate for an average time of 35 minutes, and it is therefore tempting to conclude that this is the optimal time needed for copulation. But how to test this? Parker's solution was to plot the proportion of eggs fertilized against the combined search, guard, and copulation times. He thereby found the theoretical optimum to be 41 minutes which is very close to the observed 35 minutes (Sober, Biology 133-4).

Similar measures can be adapted to study human behavior. For example, Wilson and Daly (Homicide) suggest that mothers who are selective in raising their apparently healthy infants and neglecting or even killing less healthy infants will have a higher rate of fitness compared to those mothers who do not discriminate between healthy and unhealthy offspring. This is because mothers who invest in unhealthy offspring (who, in the EEA, would have been far more likely to die anyway) will typically have fewer resources to invest in healthy offspring, thereby compromising those offspring. Daly and Wilson offer cross-cultural data which correlate infanticide with heath and other threats to infant heath (maternal health, presence of a supportive father, etc.) to suggest that maternal investment contributes strongly to fitness and that its sensitivity to variables is therefore likely an adaptation (Homicide, 37-59).

However, this thesis, though plausible, is far more tenable if it can be shown that mothers indeed invest differentially in healthy and unhealthy infants. To this end, Janet Mann conducted a three year study of seven pairs of Preterm identical

twins, measuring several forms of maternal response to each infant over a three year period. As Mann predicted, mothers "demonstrated distinct behavioral preferences for their healthier twin infant (Mann "Nurturance or Negligence", 386)" even though the presence of observers in their homes might be expected to influence mothers to "perform" for the observers by attending to each infant equally. Mann suggest that mothers may not even be conscious of their healthy child preference and are relying on a "healthy infant template" which directs their behavior (385-7).

I should note here that I have significantly simplified my discussion of both of these examples because I am not interested here in defending the theses which Parker and Mann offer, but only in showing that adaptive stories need not be treated as mere speculation since in many cases suitable tests can be devised which confirm their predictions.[100]

[100] As a postscript to this section, I add this thought: A possible and perhaps worrisome implication of these consideration is that, while genetic determinism may not be true, the combined forces of biology and the environment may in themselves be deterministic, in the sense that any human action may be fully deduced from the appropriate causal laws and knowledge of prior states of the universe. Does human moral responsibility disappear in such a case? I don't think so. Even if it did, that fact would not count against EP, since the existence of free will is not a metaphysical given, but a question to be settled by empirical means and thus subject to empirical refutation. Consequently, it is far more profitable to investigate which forces shape our decisions (to whatever degree they do shape our decisions) than to speculate on the existence of a faculty of free will. Discovering these forces will not, as some seem to fear, turn us into fatalistic robots whose actions are predestined and unstoppable. Rather, learning the springs of our hopes and fears and desires will help us understand and control them. Knowing ourselves entails understanding all the forces that shape our actions, whether these are due to environment, genes, or some mysterious "self" that is the product of neither. And if free will does exist, it probably exists as soft determinists describes it - as the absence of internal or external constraints to action, and not as a force outside of material determinism(see Daniel Dennett's Elbow Room for such an account).

The Politics of The Politics of Evolutionary Psychology

> We are the descendants of *Homo erectus*, we are told by our wise men, the anthropologists. (Could it be that men have a thing about uprightness?)
>
> (Baier 317)

Perhaps the most problematic and most frequently repeated objection to evolutionary psychology is the claim that it is not really an objective scientific view of who we are, but a value-laden, politically motivated, ideology. There is an awful lot to unpack in this statement, and it is difficult to know just where to start.

One way of expressing this complaint is this: unlike other theories of human psychology such as Hegel's master-slave dialectic or Marx's theory of false consciousness, evolutionary psychology is (or purports to be) a product of science, and therefore we need some proof that science is in fact objective before we can accept evolutionary psychology. Presumably this distinction rests on the fact that the former theories, beings products of nineteenth century idealist metaphysics, simply do not fall under under the rubric of theories that can be judged as objective or otherwise.

What then would it mean for science to be objective? Scientists who are objective will be those who employ rational methods of investigation, who are guided by the facts, and not by their non-epistemic values, and who take appropriate measures to prevent those values from affecting their conclusions. But plainly not all scientists are successful in this task, and this suggests that science cannot be objective. Critics outside science (e.g., Code 27, Harding 49) object that scientists believe in the "myth of objectivity," falsely believing themselves to be

objective and falsely presenting their work as if it were the product of objective and impartial, disinterested investigators.

But Douglas Futyama, in his much-cited Science on Trial, argues that this widely-held public perception of science is itself a myth which scientists themselves do not necessarily accept:

> In fact, scientists are just as human as anyone else. They believe that one or another hypothesis is most likely to be true, and they engage in sometimes bitter battles to defend their ideas. Scientists' beliefs are also shaped by their political, social, and religious environment. It is undoubtedly true that Darwin and Wallace were led to the idea of natural selection because the English economic system of the day put an emphasis on competition, free enterprise, and economic progress .... Thus the common image of scientists as abstracted, unbiased, detached intellects has no foundation in reality. Scientists are often highly opinionated, even in the face of contrary evidence, and they are often not particularly intelligent, either. The spectrum of scientists, as of any other group of people, runs from the brilliant to the fairly stupid. Almost every scientist has made more than one asinine statement in the course of his or her career, and some of them habitually. (164)

Moreover, since 1989, the Committee On Science, Engineering, and Public Policy of the National Academy of Sciences, the National Academy of Engineering and the Institute of Medicine have widely distributed entitled "On Being a Scientist" to students entering the sciences and engineering. This booklet (also available in an on-line version) and subsequent initiatives by various national bodies, stress the ways in which scientists need to recognize how values can and do affect

research in unacceptable ways. The claim that scientists are not objective is no longer a revelation of the marginalized; it is closer to the truth to say it is an institutionalized boilerplate assumption.

Science, however, is not merely a concatenation of scientists; it is a collective and shared practice and process which is (ideally) structured in such a way that that objective criteria will prevail over bias and error. Longino, for example, specifies how such a process should be structured:

> Scientific communities will be objective to the degree that they satisfy four criteria necessary for achieving the transformative dimension of critical discourse: (1) there must be recognized avenues for the criticism of evidence, of methods, and of assumptions and reasoning; (2) there must exist shared standards that critics can invoke; (3) the community as a whole must be responsible to such criticism: (4) intellectual authority must be shared equally among all qualified practitioners (cited in Goldman 78)

This is, of course, just one recipe for constructing an objective science, but the claim is that where this (or some other similar set of necessary and sufficient conditions) obtains, scientists will create theories that are themselves objective in that they describe the world as it really is, are consistent with the facts, are intersubjectively verifiable, etc.

But there are many well-formulated and widely-respected views which suggest that science does not in fact satisfy Longino's criteria (or any other similar criteria designed to produce objective knowledge), and that, even if it did, it would not provide objective knowledge of the world. These various views (summarized from Goldman 7-40) hold that: truth is just a matter of what social negotiation

allows (Rorty); truth is just the product of scientific negotiating which constructs a fictive reality (Latour and Woolgar); it is language which determines the contents of our beliefs (Whorf-Sapir) and people create truths through the use of language; objective, transcendent truth is in principle unattainable by humans; truth can only be evaluated within some language games and there is therefore no such thing as epistemic privilege and no neutral, non-cultural way of deciding between competing theories of the truth, that truth is merely an instrument of domination (Foucault); and, finally, all truth claims are ineradicably tainted by biases of one sort or another.

Given these criticisms, the onus is clearly on the defender of scientific objectivity to show why each and every one is false. If this cannot be done (and critics think it cannot), then all claims arising in evolutionary psychology must be embargoed from philosophical discourse.

Of course, to answer all these charges, and all their variants, would take us an impossible distance from the object of our study here, so I suspect that I may beg off this task on the grounds that space does permit me to do so. And, as I've noted above, my purpose here is only to show that evolutionary psychology is at least as plausible as most other projects within evolutionary theory. If, because of the skeptical claims made above, you harbor serious doubts about the truth of evolutionary theory, what follows may therefore no more compelling than any other bit of evolutionary theory. I have also hinted above that Hegelian and Marxist theses about human psychology do not seem open to the charge of nonobjectivity. But why should this be? In part, it is because science aspires to objectivity and because the nature of its techniques and the objects of investigation are such that cases of bias and error are often acutely visible.

Western philosophy has held a long and enduring relationship with psychology practically since its inception. Although psychology "moved out" on its own a century ago, we frequently treat psychological claims made by philosophers (e.g., Plato's tripartite soul, Rousseau's noble savage, Hobbes' account of man in the state of nature, Locke's theory of the *tabula rasa*, Mill's argument defending the value of higher goods, Nietzsche's *ressentiment*, etc.) more as integral parts of philosophical systems and less explicitly as generalizations about humans which may or may not have been rendered obsolete by more recent and philosophically unnoticed advances in psychology and other social sciences.

Yet philosophy - especially political and moral philosophy - continues to rely heavily on this sort of armchair psychology, and perhaps often without the conscious realization that the line between metaphysical speculation and empirical claim-making since has been crossed. Certainly moral and political philosophy could not survive without <u>some</u> account of human nature which can usefully support our claims about what people want, how they are likely to react to difference in political structures, what effects punishment is likely to have on them, how much heroism they are capable of, etc. And (as the tradition implicitly approves) where science or religion cannot provide us with a detailed and indubitable account of human nature, we are surely entitled to invent one. After all, political problems are here and now, and to be agnostic in the face of great wrongs because one is uncertain about the actual shape of human nature seems worse than unwise and imprudent.

So, as philosophers, we can construct our own accounts of human nature, so long as we are modest in our assumptions. Or perhaps even this limit does not apply. Carole Pateman, for example, has "retrieved" an account of the original social

contract which, she claims, was founded on the resolution of a putative primordial Oedipal complex (104ff). In saying this, is Pateman committed to overcoming the many objections to Freudian theory (as well as the set of deeply skeptical objections I've noted above which are directed at science in general)? Or is she simply telling a story which is plausible so long as it concurs with common sense and seems to explain our current political arrangements? In any event, it should be clear that evolutionary psychology is no less illegitimate than any other psychological heuristic (such as Freudianism) within philosophy. Insofar as it coheres with, and is supported by, the great weight of evidence for evolutionary theory, this counts greatly in its favor. So, given these considerations, I propose to skirt global objections to science's objectivity entirely. In place of this, I want to discuss the much thornier and more pertinent problem of politics within the field of evolution itself, because this seems to be a major barrier to the wider acceptance of EP.

Birke and Silvertown's treatment is effectively terse and representative of the genre:

> The life sciences are not value free: they make social and political assumptions. In recent years, biologically-based arguments have been used to bolster a variety of right-wing policies and prejudices. Feminism is opposed on the grounds that male domination is a product of male biology; racism is justified with the argument that we have a dislike of strangers wired into our genes; elitism on the grounds that differences in intelligence are innate; the arms race on the basis that competition is natural and desirable. (rear cover)

Rose et al (*passim*), Lewontin (Ideology 20-37), Tuana (621-3), Code (128-29), the

BSSRS Sociobiology Group (110-35) all tell parts of the same story: The notion of innate determinism predates genetic theory and even predates evolutionary theory: A series of research programs (all deeply flawed, and all politically motivated) sought and and found evidence of innate causes that explained putative differences between whites and other races, between men and women, between the rich and the poor, between "Aryans" and Jews, between law-abiders and criminals, between the sane and the insane, and so on. In each case, since the cause was innate, it was deemed fixed, a result of the iron hand of nature. And in each case, oppressive ideologies (racism, sexism, Nazism, eugenics, Social Darwinism, discriminatory immigration policies, etc.) were erected on these "biological" principles - but of course, those biological principles were fabricated for the explicit purpose of legitimizing those very ideologies. Since modern evolutionary psychology is an outgrowth of these earlier discredited theories (Val Dusek calls evolutionary psychology "sociobiology sanitized" - and of course sociobiology itself is simply Social Darwinism sanitized), contemporary practitioners of EP are entrenched in the same ideological mire.

I think the crucial question here is whether or not contemporary evolutionary psychologists hold political views which so bias their research that we should not believe them, and it is this point on which I want to offer the most sustained argumentation. But first I should make four quick points to clear up some confusions which seem to arise frequently.

The first point is that, as a piece of intellectual history, the account that Rose, Lewontin, and others offer is all right as far as it goes - but it does not go far enough. Much like religion (with which it is often compared) the theory of evolution is a broad and immensely fecund metanarrative consistent with many diverse political interpretations. Just as we do not expect co-religionists to hold

identical political positions, nor should we expect evolutionists to converge in theirs. And, as Michael Ruse points out, the historical record bears this out, showing that evolutionists display a variety of political views. Among them: Social Darwinism conceived of as offering the disadvantaged a way to improve themselves (Andrew Carnegie); anarchism, mutual aid and cooperation (Kropotkin); female control of sexual selection (Alfred Lord Wallace); activism in defense of biodiversity (E. O. Wilson Biodiversity); cooperation (John Rawls Theory, 502-4); utilitarianism (Peter Singer); neo-Jungian feminism (Genia Pauli Haddon "Yang-Femininity). And so on. The point is not that any or all of these are legitimately supported by evolutionary theory: the point is that one is in no way committed to any right-wing or oppressive doctrine simply in virtue of understanding humans within a Darwinian framework, and the historical record proves this. Ruse concludes, "it is historically inaccurate to present evolutionary theorizing past and present as one story of morally and culturally offensive proselytizing (Naturalism 211)."

Second, this theory emphasizes the evils of building political empires on theories of biological determinism - we have only to look at Nazi Germany to see where that will lead us. But the evils of social determinism are no less appalling - look at the disasters wrought by social engineering in the former Soviet Union, China, and Cambodia. Given the fact that three of the ablest and most respected critics of sociobiology and evolutionary psychology - Stephen Jay Gould, Stephen Rose, and Richard Lewontin - are all dedicated Marxists who are explicit in explaining how their politics has influenced their biological theories, this point is doubly acute. Nonetheless, it is as unfair, unhelpful, and mischievous to accuse them of aiding and abetting genocide via their application of doctrinaire Marxism to biology, as it is to accuse Richard Dawkins and E. O. Wilson of providing aid and

comfort to neo-Nazis.

Third, as critics all admit, all these deeply undesirable biologically based ideologies rely on the assumption of hard biological determinism. But, as I've shown, there is scant evidence to show that contemporary evolutionists are committed to hard biological determinism.

Finally, it is tempting to reject evolutionary psychology simply because it appears to have undesirable political consequences. I think Charlotte Bunch does precisely this in this passage:

> Female subordination runs so deep that it is still viewed as inevitable or natural, rather than seen as a politically constructed reality maintained by patriarchal interests, ideology, and institutions. But I do not believe that male violation of women is inevitable or natural. Such a belief requires a narrow and pessimistic view of men. If violence and domination are understood as a politically constructed reality, it is possible to imagine deconstructing that system and building more just interactions between the sexes. (65)

Of course, there is no reason whatever to think that because some psychological disposition is "natural" that it is therefore inevitable. But it is an even greater error to make a sweeping generalization about millions of people (even one that is quite likely to be true) simply on the grounds that if it were true, it would make a very desirable political outcome possible. There is a very powerful and well-known argument which explains exactly why such inferences cannot be made. As David Hume pointed out, we cannot infer that because some state of affairs is natural, that it is therefore morally desirable (Treatise, III.i.1). And the

contrapositive is equally true: we cannot infer that because some state of affairs is morally undesirable, that it is therefore unnatural.

We now need to consider whether the likely level of bias among evolutionary psychologists counts as a good reason to disbelieve their claims. The literature claiming bias is ugly, anecdotal, inconclusive, and probably not worth our consideration. I don't think, in the end, that it will reward us to consider whether Derek Freeman dislikes naked genitals on public statues, or if Margaret Mead was a homophobic militarist, or whether Stephen Pinker revealed his deepest political commitments when he announced that he didn't really like John Lennon's song "Imagine" (Dusek, "Sociobiology"). And it will be even less useful to speculate about what effect these putative beliefs had on their work. In fact, I think that once an allegation of bias - linking Dawkins' work to the National Front, for example - is raised, no amount of apologia ("But some of my best friends are black") will quell the suspicions. So I propose instead to argue that genetic arguments based on claims of bias carry much less weight than many people suppose, and that they may even decrease our epistemic success.[101]

Why Bias Isn't so Bad

One reason why we might aspire to a value-free science is that some values - call them "biases" - will unacceptably influence the way scientists frame scientific questions, select, interpret, and reject data, devise methodologies, and formulate explanations. Since theories are always underdetermined by observation,

[101] I thank Paul Viminitz, Oliver Schulte, Elliot Sober, Wesley Salmon, David Hitchcock, and the members of the Department of Philosophy of the University of Lethbridge for helpful comments on earlier drafts of this section. I am also indebted to Glenn Parsons for suggesting that Bayes' Rule might usefully unpack this problem.

observer bias will fill the epistemic gap and create observer-relative theories. In short, bias can taint an entire scientific enterprise with nonobjectivity. Following Robert Nozick, let us call this the Contamination Thesis ("Invariance" 34). It is now tempting to suggest that that if some theory Q is tainted by bias, that fact will count as a reason to think that Q is false. If we combine the Contamination Thesis with the further claims that (1) bias is endemic to some research tradition and (2) that this bias is the chief causal force behind theory construction, promulgation, and acceptance within that tradition, we now have a powerful conceptual tool with which to reject entire research traditions.

However, the Contamination Thesis is not sufficient to establish that a theory is false. Practical logic warns us that genetic arguments such as these are invalid and fallacious since even apparently disreputable origins can yield true claims.[102] But there are at least three suggestions that this is not the whole story.

1. Lorraine Code points out that there is a curious asymmetry between the way we consider appeals to authority and and the way we treat genetic arguments (27). To paraphrase Code slightly, if an unbiased and credible authority utters some claim P, we typically count this as a good probabilistic reason to think that P is true. By parity of reason then, if a biased and therefore non-credible source utters Q, this ought to be a good probabilistic reason to think that Q is false. In other words, genetic arguments are (or ought to be) the epistemic mirrors of appeals to

[102] Copi and Burgess-Jackson, 121-2.I attach no great weight to the concept of "truth" here, nor need I claim that science ever makes any true claims. Those who are skeptical of truth can simply consider the word a bit of shorthand for whatever degree of epistemic approval they would confer on statements of the sort "there are at least three people in this room" or "there is at least one scientist who is biased."

authority.[103]

2. Alvin Goldman, following Wesley Salmon, also argues for parallelism between appeals to authority and genetic arguments if those arguments are based on an agent's epistemic history. Suppose an agent makes some set of claims about subject S and the great majority of these claims prove false. If the agent makes some further claim Q about S, Goldman claims it is then simply a matter of "proper induction" to conclude that Q is false as well (152-3).

3. Elliot Sober has argued that insofar as an agent's belief is independent of the truth of the belief, it is for that reason likely to be false. Given these considerations, it seems that a probabilistic version of the genetic argument may be acceptable where a deductive version is not.

I contend to the contrary that both the "hard" (deductive) and "soft" (probabilistic) variants of the genetic argument are flawed. Except for some exceedingly rare cases which I'll discuss later, evidence of human bias is never a good reason to think a claim is false. I shall use Bayes' Rule and intuitively acceptable arguments to show that Sober's argument is mistaken and shall offer some normative suggestions about the epistemic value of genetic arguments.

[103] I don't read Code as claiming that all appeals to authority and all genetic arguments are strong arguments. After all, people frequently ascribe degrees of epistemic authority based on mistaken beliefs or irrelevant criteria. I think what Code is pointing at is that since at least some appeals to authority have probabilistic merit, so too do some genetic arguments.

Sober's Probabilistic Genetic Argument

Sober argues that even though deductive forms of the genetic argument are indeed invalid, this point has been over-interpreted.[104] There are, he argues, perfectly respectable probabilistic versions of the genetic argument.[105] Sober offers this thought experiment:

> Suppose I walk into my introduction to philosophy class one day with the idea that I will decide how many people are in the room by drawing a slip of paper from an urn. In the urn are a hundred such slips, each with a different number written on it. I reach in the urn, draw a slip that says "78," and announce that I believe that exactly 78 people are present. (206)

Since Sober's belief is almost certainly incorrect, Sober thinks we can construct the following genetic argument:

> (1) Sober decided that there were 78 people in the room by drawing the number 78 at random from an urn.
>
> p ==========:==========
>
> It isn't true that there are 78 people in the room.

The "p" and double line indicate that the argument is non-deductive and that the premise confers probability p on the conclusion. Sober contends that p is high

[104] Sober Biology, 206; Point of View, 105. Sober's discussion of the genetic argument in Philosophy of Biology and From a Biological Point of View are almost identical. I have used the former account throughout.

[105] Sober's intended target is a subjectivist argument that denies the validity of ethical statements by attacking their evolutionary and social origins, but the form of his argument allows it to be employed much more widely.

and that this is "a perfectly sensible genetic argument" in which "the conclusion is justified *because* of the process that led me to this belief (206, emphasis added)" even though "*what caused me to reach the belief had nothing whatever to do with whether the belief is true* (207, emphasis in original)." By way of contrast, if Sober's alter ego Rebos carefully counts all the people in the class and consequently believes there are 104 people present, we have good probabilistic grounds to think that Rebos is right, because she arrived at her belief in a respectable way. Sober's moral is this: where an independence relation holds between a belief's cause and the truth of the belief, the belief is likely false. Contrariwise, if there is a dependence relation between the belief's cause and its truth, the belief is likely to be true. So this example proves that genetic arguments can offer probabilistic grounds to think some claims are false (207). This interpretation looks intuitively compelling, but I think it's incorrect, and Bayes' Rule demonstrates this conclusively. According to the simplified version of Bayes' Rule:

6. $P(Q \mid R) = (P(Q) \times P(R \mid Q)) / P(R)$

where:

7. Q = There are exactly 78 people in the class.
8. P(Q) = The probability that (Q)

9. R = Sober randomly draws the number 78.

10. P(R) = The probability that (R) = 1/100

11. P (Q | R) = The probability that there are exactly 78 people people in the class conditional on the fact that Sober randomly draws the number 78.

12. P (R | Q) = The likelihood that Sober draws 78 conditional on the fact that there are exactly 78 people in the class. R and Q are independent, so P (R) = 1/100.

Substituting these values in (6) yields

13. P(Q | R) = (P(Q) X 1/100)/(1/100)

which simplifies to:

14. P(Q | R) = P(Q)

In other words, the probability of there being 78 people in the room given that Sober drew 78 from the urn is exactly equal to the prior probability that there are 78 people in the room.[106] Therefore Sober's conclusion that there are exactly 78 people in the room is probably false just in case we think P (Q) is small. But Sober's argument (1) offers no evidence whatsoever that P (Q) is small. This suggests that the argument relies crucially on a suppressed premise:

(5) Sober decided that there were 78 people in the room by drawing the number 78 at random from an urn.

[106] The same conclusion can be reached directly via the definition of probabilistic independence, without the need to appeal to Bayes' Theorem.

[University classes rarely contain exactly 78 people.][107]

p ==========

It isn't true that there are 78 people in the room.

Making this suppressed premise explicit shows that it, and not Sober's independence thesis, is in fact doing all the evidential work. If you doubt this, consider these two variants of Sober's argument:

(6) Sober decided that there were 78 people in the room by drawing the number 77 at random from an urn.

[University classes rarely contain exactly 78 people.]

p ==========

It isn't true that there are 78 people in the room.

(7) Sober decided that there were 78 people in the room by drawing the number 78 at random from an urn.

[Universities rigidly enforce rules requiring there to be exactly 78 people in every class.]

p ==========

It is **true** that there are 78 people in the room.

[107] Jennifer Welchman and David Hitchcock (p.c., 2001) have both challenged the need for this particular premise, suggesting that it could be replaced by one that states that random choices rarely correctly predict class sizes. Hitchcock further objects that my proposed premise renders the first premise irrelevant. But as the mythical Microsoft helpline technician is supposed to have said, "That's not a bug, that's a feature." My motivation for inserting this premise is precisely to render the new information about the random draw irrelevant. My application of Bayes demonstrates that the posterior probability is exactly equal to the prior probability. Therefore, the new information is irrelevant and the argument, if it is to conform with Bayesian constraints, must contain all the information that defined the prior probability. And to be more precise, my premise (2) ought to note that not all class sizes are equiprobable. The fact that the alternative premise (2) cannot yield a conclusion which is sensitive to the differential probabilities of different class sizes should also alert us to the fact that the posterior probability that it yields cannot be equal to the prior probability. Note that adding the revised premise (2) to argument (7) makes this explicit.

Now let me make the disagreement between Sober and me as explicit as possible. Sober thinks that:

> (8) "drawing 78 at random" and believing that "there are 78 people in the class" are two <u>independent</u> states of affairs.
>
> ---
>
> (1) is a "convincing" argument (207).

While I contend that

> (9) "drawing 78 at random" and believing that "there are 78 people in the class" are two <u>independent</u> states of affairs.
>
> ---
>
> (1) is a weak argument.

I want now to diagnose just <u>why</u> we disagree. Consider for a moment the following matrix:

	INDEPENDENT BELIEF FORMATION: Sober draws the number 78 from an urn and consequently ...	DEPENDENT BELIEF FORMATION: Rebos carefully counts the people in the room and consequently ...
YIELDS TRUE BELIEF (PROBABLY)	POINTLESS RANDOM CHOICE: Sober believes there are NOT 78 people in the room.	EMPIRICISM: Rebos believes there are 104 people in the room.
YIELDS FALSE BELIEF (PROBABLY)	RANDOM CHOICE: Sober believes there are 78 people in the room.	PERVERSE EMPIRICISM: Rebos believes there are NOT 104 people in the room.

Table 4: Independence and Belief Formation

We can now see why the arguments offered by Sober, Code, and Goldman all fail to justify probabilistic genetic arguments. Sober only considers the beliefs represented by the lower left hand and upper right hand boxes and this, I suggest, is why he thinks that a belief's plausibility is linked to its dependence. But this inference is mistaken. In the first column, the cause of both of Sober's beliefs is independent of the facts. But the belief generated by Pointless Random Choice in the upper square is almost certainly true. Hence it is false to say that the independence relation cannot reliably produce true beliefs. And the belief produced by Perverse Empiricism is almost d certainly false. A Perverse Empiricist believes some claim Q iff she has carefully investigated Q and found compelling evidence that Q is false. Hence her beliefs are dependent on the truth, but, so to speak, inversely so. Hence it false to say that beliefs which are dependent on the truth are likely to be true. These two examples prove that dependence is neither necessary nor sufficient for true belief. We also need to recognize that Sober's example also fails (no doubt for the sake of lucid exposition) to model real world belief formation procedures where degrees of

epistemic in/dependence are far more difficult to ascertain and where prior probabilities may not be so obviously minuscule. If my line of reasoning is correct, Sober (as he now agrees) has given us no reason to think that the origins of a belief will count as reasons to think that it is false - even where those beliefs are reached on the basis of random choice.

Earlier, I interpreted Lorraine Code as arguing that genetic arguments should be the epistemic mirrors of appeals to authority. The lower right hand corner represents beliefs formed by Perverse Empiricism, and this seems to fit Code's position. This is the only case in which a genetic argument has any force. Precisely because we know the Perverse Empiricist's epistemic practices reliably create false beliefs, appeals to those practices count legitimately as a reason to think her claims false. So, properly considered, the argument from authority and the genetic argument are indeed analogous. And it might be tempting to think that Perverse Empiricism models real-world cases of scientific bias. But I think this assumption is seriously mistaken and likely to lead to great confusion. Consider a completely biased claim-maker who will assert Q iff Q supports his bias. Call this condition "complete bias."[108] A completely biased claim maker will accept all and only those data points which support his bias. Importantly, it is not the case that he rejects claims because he thinks them true, so his beliefs do not correlate negatively with the truth.

Suppose for example that our completely biased claim-maker is a judge who convicts all native suspects. Even if I have justified and irrefutable knowledge that the judge convicts all native suspects no matter whether they are guilty or

[108] It may that many people think of all cases of bias as cases of complete bias. But I think this assumption is unwarranted - it seems more likely that in most cases of bias, an investigator may select some data points because he thinks them true independently of his biases, and he may even accept some data which are contrary to his bias.

not, the fact that I know that some suspect has been convicted by the judge gives me no reason to think that the suspect is innocent. The only reason to think the defendant is innocent is whatever knowledge I possess about the prior probability of the innocence of native suspects.

Now consider by way of contrast the Perfectly Perverse Empiricist. This claim maker accepts only that data which he knows to be false. Substituted for our completely biased judge, he would convict all and only those native suspects who are innocent. In this case, knowing that the Perfectly Perverse judge had convicted a native suspect would be very good reason to think the suspect was indeed innocent. But I contend that this is not the case in real world instances of bias. Biased claim makers do not typically reject claims because they are true; they reject them because they do not support their biases. The typical biased claim maker is quite happy to accept a claim that also happens to be true, but the Perverse Empiricist is not: he is, in a sense, "allergic" to holding true beliefs. And this distinction marks a fundamental difference between the two. Because Perverse Empiricism is pathological, rare, and largely irrelevant to scientific study, I will henceforth disregard it.[109] This realization shows that Lorraine Code's argument from analogy is of little epistemic utility.

Goldman's contention that one can make a persuasive inductive argument from an agent's past false claims fares no better. Suppose Jones makes a set of false claims {A ...P} about subject S. If Jones further asserts claim Q about S, isn't her espousal of the earlier set of false claims a good reason to think that Q is also

[109] Measuring devices might count as a counterexample to this generalization, since they may reliably produce inaccurate results (too high or too low.) And some might think this a type of bias. But this will not affect my comments about human bias.

false?[110] No. If there is some direct logical or evidential link between {A ...P} and Q such that the falsity of the former is evidence for the falsity of Q, then you have good reason to think Q is false independently of Jones's actually asserting Q.[111] On the other hand, if Q is logically independent of the former discredited claims, then its probability cannot be less than its probability prior to Jones asserting it, and this follows for precisely the same reasons I adumbrated above against Sober's example. In neither case is the shared origin of claims {A...P} and Q relevant to our evaluation of Q's truth value. No matter how poorly my previous weather predictions have been in the past, my dismal record to date is no reason whatsoever to have any skepticism about my current claim that it it is likely to be above minus 60 Celsius tomorrow in Edmonton.

[110] Even determined liars must admit to self-evident truth in order to profit by their crimes. For example, it is a standard ploy in counterintelligence to have double agents deliver false information to one's enemy, that the enemy will use to its own misfortune. However, in order to convince the enemy that the disinformation is in fact credible, it is necessary to also divulge some secrets that are true, important, and perhaps even harmful to one's own interests. The Perverse Empiricist, by definition, does not do this and it is therefore a mistake to see her as a sort of liar.

[111] I'm not sure it is possible to exhaustively list how the falsity of {A...P} could be evidence for Q's falsity. If Q entails the former set, then this would be sufficient.

Evidence of bias is thus never, all by itself, a reason to disbelieve any claim Q. And if the prior probability of Q is high, it may not even be reason to be agnostic about Q.[112] Nonetheless, it is straightforwardly false to assert, as Annette Baier does, that arguments against genetic arguments require us to "ignore" Q's origins as irrelevant (325). After all, in many cases, knowing that Q comes from unsavory origins will be good reason not to increase one's degree of assent. But this will only be the case where a testifier is strongly biased towards Q. Where an agent is only weakly biased towards Q, it may be rational to increase one's belief in Q even though one knows that the testimony for Q is biased. Consider, for example the testimony of two chicken sexers. One is strongly male-biased- she identifies all chickens as male, whether they are or not. In this case, we should retain our prior belief that any given chicken has a 50% probability of being male. The other is only weakly pro-male in her assessments - she correctly identifies all male chicks as male, and misidentifies only 1% of female chicks as female. It is therefore rational to believe the testimony of the latter (and to believe it to a high degree) - even though we know her testimony to be biased. But suppose we do not know if the chicken sexer is biased or not and that we cannot even assign a probability to the possibility of his being biased? If we know that the only bias the chicken sexer is likely to have is a pro-male bias, it may still be rational to revise our estimates of a given chicken's sex slightly based on his

[112] For my purposes here, I define "agnostic" as having no reason to prefer one truth value over another for some claim. Formally, the agent might say she assigns a probability of 0.5 to that claim. But it probably more common that she is unwilling to be so precise. In these cases, it is sufficient for my purposes that the agent be much more willing to accept a probability close to 0.5 than she is to accept a probability close to 1 or 0. Now consider the case in which exactly one of four theories [A, B, C, D] is true, and one has no reason to prefer any of them (i.e., P(A) = P(B) = P(C) = P(D)). The prior probability of not-C is therefore 0.75. Later evidence might increase one's belief in not-C, but if this evidence turns out to be biased, one is still justified in assigning not-C a relatively high level of probability (0.75).

pronouncements. But if we do not know if the chicken sexer is strongly or weakly female-biased, strongly or weakly male-biased, or completely unbiased, it seems the most prudent course is to ignore her judgments altogether.

So a rational agent who desires to maximize her quotient of true (or justified) beliefs over false beliefs will be acting in an epistemically responsible manner if she takes credible accusations of bias seriously. I offer some modest suggestions about how she should do this.

First, it is tempting to believe that if a scientific claim Q has a low prior probability, one can therefore infer the likely existence and influence of bias in its creation. Given this, one can next deduce the nature of that bias from the content of Q itself, and this inference can then be used to erect a genetic argument against the veracity of Q. (Lorraine Code's dialogical epistemology apparently licenses this methodology.[113]) This practice is, however, deeply flawed for several reasons. First, the existence of bias or error will not be revealed in the prior probability of Q itself, but in the conditional prior probability of an agent's asserting Q given that Q is in fact true (Nozick, Rationality 101).[114] Let me explain this a bit. If, for example, one knows that Q is false, one might guess which biases might have led a researcher to espouse it. Contrariwise, if one knows that a researcher has a given set of nonscientific commitments, one might guess how those

[113] Lorraine Code employs this form of argument against Philippe Rushton, who has argued (notoriously) that there is an interracial inverse correlation between penis size and intelligence. Code contends that since Rushton could not have found his data "by coincidence," he must therefore have been driven by some right-wing agenda. Code proceeds to lay out in detail what Rushton's politics must be, and then suggests that the existence of these politics constitutes a probabilistic reason to reject Rushton's work (28-9). This tempting conclusion, however, rests on a false dilemma (between discovery by coincidence and discovery motivated by right-wing bias), since there are any number of motivations that could have informed Rushton's research.

[114] For example, there may be a very low prior probability that I win a lottery (say, 1 in one million). But it is not rational to therefore conclude that any report that I have won a lottery is unlikely to be true. What is relevant is whether the testifier could know that I have won the lottery, given that I have indeed won the lottery.

commitments would affect her work. But both of these approaches are very dodgy enterprises, since there are any number of nonscientific considerations which might have motivated a researcher, and since a researcher's political and moral commitments do not exert a deterministic influence on her scientific claims. After all, people frequently do arrive at counterattitudinal beliefs (Goldman, 236). Now consider the case in which one knows neither that Q is false nor that bias played a role in Q's construction. In this case, to infer the existence of bias from the content of Q and to then argue that that bias now counts against the truth of Q is surely to build epistemic castles in the air. And, as the Bayesian argument above shows, erecting probabilistic arguments against Q which are based on Q's own low prior probability will lead to double discounting. So, to avoid these evils, claims that a researcher is biased should be based on independent evidence about the researcher's nonscientific commitments.

Some science critics (Rose et al, 8; for example) have assumed that this measure is sufficient all by itself: prove that a scientist has a given political commitment and you've proved that it also adversely affects her research. But whether this is so is an empirical question, and must be settled by empirical means. Alvin Goldman argues that many case studies on scientific bias are hampered by several flaws. First, studies which show that political interests are coincident with claim Q cannot, by their very nature, establish the counterfactual condition that had those political facts not obtained, the claim Q would not have been made. Such case studies therefore cannot establish the causal efficacy of politics on the development of Q. And even where they do, they are less persuasive in explaining Q's continued acceptance. Finally, Goldman suspects, many case studies are not undertaken on a random or representative set of scientific

episodes, but are handpicked to prove the very points which science critics wish to make (37-40).[115] Consequently, these case studies cannot be used to license generalizations about the effect of bias across all science. Given all this, it seems that the best way to conclusively prove that bias has led a researcher astray is to show first that she did go astray, and then to show that bias was the cause. But where this can be done, one of course no longer needs a genetic argument.

Furthermore, Robert Nozick has argued that no factor is intrinsically biasing and that whether or not a factor biases epistemic products depends crucially on the process in which it occurs. For example, although jurors are supposed to be unbiased, it may well be that a jury with two biased and opposed jurors will more frequently arrive at the truth - and this, again, is an empirical question ("Invariance", 33-4). It is even plausible that the presence of political bias can in some cases increase a researcher's credibility. If, for example, feminist anthropologists have political interests in discovering evidence of ancestral matriarchies, then when feminist anthropologists such as Pam Bamberger and Sherry Ortner fail to find any such evidence, this is particularly persuasive in showing that matriarchies did not in fact exist (example adapted from D. Brown Human Universals, 52).

The Outgroup Bias Effect

All this aside, one might still think that an awareness of bias cannot help but improve one's critical objectivity, especially when one cannot assign any prior probability to Q. My final point suggests that this may not be so. To see why,

[115] That is, critics of scientific bias may themselves be displaying bias in their selection of data to support their claims about the dangers of bias.

notice first that attributions of bias are frequently made in the third person. You and I, gentle reader, have our commitments. They have biases. We have intuitions. They have prejudices. Notice also that accusations of bias are commonly made on the basis of some difference between us and them. That is, if I warn you about Jones' dualist, antifeminist, or reductionist bias, I typically do so in the belief that you and I are not dualists, antifeminists, or reductionists. So when I ascribe bias to some third party, this accusation will frequently elicit in my listener or reader what social psychologists call ingroup/outgroup bias. This well-known and pronounced bias displays three relevant features:

1. **The Minimal Group Paradigm:** The listener will display bias against the outgroup even when she knows the differences between groups are minimal. Investigators have found that members of one group will discriminate against another group even when they know that the groups have been divided on the basis of such irrelevant criteria as a coin toss or differing preferences in modern art.

2. **Outgroup Derogation:** The listener will tend to favor the ingroup over the outgroup and will attribute more negative attributes to the outgroup. Sandra Harding, in a striking example, attributes the numerous failures of mainstream science and the many epistemic successes of marginalized knowers to the fact that marginalized knowers can somehow throw off their "covers and blinders" and thereby understand the world "how in fact it is" while scientists are "destined" to study not nature itself, but only "socially constituted

objects (54, 64)."[116]

3. **Stereotyping**: The listener will tend to believe that there is far less intragroup diversity within the outgroup than within the ingroup. Val Dusek, for example, flatly asserts that "Certainly none of the evolutionary psychologists support major egalitarian change in social or gender arrangements."

Ingroup/outgroup bias is a robust, widespread, and highly confirmed phenomenon and it is almost impossible to overestimate its effect on our epistemic practices.[117] If this model is correct, the mere suggestion that some theory Q serves some pernicious social function, or that Q is situated within some noxious political nexus, or that Q is the product of an oppressive power structure, et cetera, will thereby condemn Q as the doctrine of outgroup members whose beliefs have irredeemably contaminated their scientific understanding. Notice that the very rhetoric of "contamination" supports the gratuitous assumption that it is only ideologies which we find repugnant that could contaminate science. If, for example, one is a liberal, one would hardly say that Jones' liberal commitments had "contaminated" her scientific views.[118] And

[116] It is important to note that Harding is not merely saying that marginalized knowers of a certain type have greater epistemic success than "malestream" scientists. Her much stronger claim seems to be that scientists cannot perceive nature as it is itself, but can rather perceive only a socially constructed simulacra. Harding's marginalized knowers, on the other hand, seem to be able to perceive noumenal reality itself. No matter how one understands social constructivism, this state of affairs does not seem to be possible, since social constructivism is not a doctrine about knowers within some particular tradition, but about all knowers. The "reality" of one group is not more to be preferred than the "reality" of another. It follows form this that self-exceptions and exceptions for politically preferred groups are not permitted. And, on my reading at least, Harding offers of no argumentation to support this distinction. See Berger's and Luckmann's Reality for extended discussion.

[117] Argyle, 173-174. Pinker, 313. Tyler et al, 2-3.

[118] Since the case of the liberal here is merely an example, a placeholder for anyone who adheres to some ideology, I need not offer a definition of "liberal".

the claim of "contamination" further implies - again gratuitously - the the degree of bias is absolute and total, where in fact only a slight bias may be present. (It is incoherent to suggest that a research program is "contaminated" with political bias and then to suggest that its testimony is still somewhat credible.) And the assumption that all bias contaminates completely invites the further gratuitous conclusion that if there is a political cause for believing Q, there is therefore no epistemic reason to believe it. Even worse, such accusations will prevent us from seriously considering Q, because serious consideration of Q implies the possibility of conversion, and if the group bias theory is correct, many of us will fear conversion to outgroup beliefs more than we fear error.

If group bias effect is universal, powerful, and anti-veritistic, nostrums counseling open-mindedness are simply not sufficient. Rather, we should take care to construct arguments in ways which do not trigger well-known epistemic failings (such as group bias) in our listeners. The naturalized epistemologist can reasonably object that group bias is an epistemic pattern which must be doing some useful work for us.[119] And my concerns about the argument from bias apply with equal force against my own appeal to group bias. That said, I am not certain that arguments which rely, even implicitly, on group bias are, on the whole, epistemically advantageous within modern science.[120]

In short, I have tried to show here that the history of evolutionary psychology and sociobiology does not support the claim that these research programs have

[119] Paul Viminitz suggests this point. p.c. 20 September 2000.

[120] Of course, given Nozick's comments about the contextuality of bias, this claim needs to be nuanced as well. A group of experts (brain surgeons, say) may be well advised - on average - to value the opinions of their colleagues over outsiders, and it may even be epistemically irresponsible for them to consider expert and non-expert views as equally meritorious. Nonetheless, this heuristic will preclude or delay adoption of outsider insights to which experts are blind. So expert outgroup bias will only be epistemically justified if the experts are confident they have a strong monopoly on discoveries.

invariably been used to support reactionary political movements. There is therefore no strong inductive case that can be made to suggest that current researchers in EP must be pursuing the same reactionary political goals their forebears supposedly pursued. Nor can arguments about the putative political consequences of EP or about scientists' moral responsibility for the political ends to which their research is employed offer any guidance as to the truth of EP itself. In fact, arguments of this sort are more likely have a negative effect on our ability to rationally appraise EP. Moreover, I have tried to show that many attempts to demonstrate bias rely on implausible assumptions, and that even where there is good evidence of bias, this will never be sufficient to think that any scientific claim is false. Finally, I have argued that accusations of bias are likely to trigger a strong outgroup bias response in the listener, and this will also decrease epistemic success.

My intent here has not been to trivialize the role of bias nor to counsel quietism. Rather, my modest suggestion is that accusations of bias may be incapable of bearing all the epistemic load which they are sometimes asked to support.

If I chance to talk a little wild, forgive me;
I had it from my father.[121]

[121] Shakespeare, Henry VII. (1613), act 1, sc. 4.

Works Cited

Adams, Robert. "The Sense of Verification: Pragmatic Commonplaces about Literary Criticism." Myth, Symbol, and Culture. Ed. Clifford Geertz. New York: W. W. Norton, 1971.

Baier, Annette C. "A Naturalist View of Persons." Moral Prejudices: Essays on Ethics. Cambridge, Mass.: Harvard UP, 1995.

Barkow, Jerome H., Leda Cosmides, and John Tooby, ed. The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford: Oxford UP, 1992.

Barron, Guillermo. Plantinga on Function and the Design Plan. Edmonton: 1994.

Bateson, Patrick. "Does Evolutionary Biology Contribute to Ethics?" Biology and Philosophy. 4 (1989): 287-301.

Bedau, Mark. "Can Biological Teleology be Naturalized?" The Journal of Philosophy. (1991): 647-55.

Berger, Peter L., and Thomas Luckmann. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. New York: Doubleday, 1966.

Berkow, Robert et al, eds. The Merck Manual of Medical Information. Whitehouse Station, N. J.: Merck, 1997.

Bigelow, John and Robert Pargetter. "Functions." The Journal of Philosophy. 84.4 (1987): 181-197.

Boden, Margaret A. "Horses of a Different Color?" Ramsey, Stich, & Rumelhardt. 3-18.

Bouchard, Thomas J. and Matthew McGue. "Genetic and Rearing Environmental Influences on Adult Personality: An Analysis of Adopted Twins Reared Apart." Journal of Personality. 58.1 (1990): 263-292.

Boorse, Christopher. "Wright on Functions." Conceptual Issues in Evolutionary Biology: An Anthology. Ed. Elliot Sober. 1st. ed. Cambridge, Mass.: MIT P, 1984. 369-85.

Brown, Donald E. Human Universals. Philadelphia: Temple UP, 1991.

Brown, Harold I. Rationality. London: Routledge, 1988.

BSSRS Sociobiology Group. "Human Sociobiology." Silvertown. 110-135.

Bunch, Charlotte, "Women's Rights as Human Rights." Applied Ethics: A Multicultural Approach. Ed. Shari Collins-Chobanian, and Kai Wong. 2nd. ed. Upper Saddle River: Prentice-Hall, 1998. 61-71.

Bunzl, Martin. "Baseball and Biology." Philosophia. 27.3-4 (1999): 575-79.

Buss, David M. "Mate Preference Mechanisms: Consequences for Partner Choice

and Intersexual Competition." Barkow et al. 249-266.

Byrne, Richard W. and Andrew Whiten, ed. Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans. Oxford: Clarendon Press, 1988.

Calvin, William H. The Ascent of Mind: Ice Age Climates and the Evolution of Intelligence. New York: Bantam, 1990.

Cann, Rebecca L., Mark Stoneking and Allan C. Wilson. "Mitochondrial DNA and Human Evolution." Nature. 325 (1 January 1987): 31-6.

Cherniak, Christopher. Minimal Rationality. Cambridge, Mass.: MIT Press, 1986.

Christensen, David. "Robert Nozick, The Nature of Rationality." Noûs. 29.2 (1995): 259-274.

Code, Lorraine. "Taking Subjectivity into Account." Feminist Epistemologies. Ed. Linda Alcoff and Elizabeth Potter. New York: Routledge, 1993. 15-48.

Cooper, Wes. "Parfit, Heroic Death, and Symbolic Utility ." Forthcoming in Journal of Social Philosophy, 2001.

Cooper, Wes and Guillermo Barron. "Buridan's Ass: A Decision Value Approach." Forthcoming in Philosophy in the Contemporary World, 2001.

Cooper, W. S. "How Evolutionary Biology Challenges the Classical Theory of

Rational Choice." Biology and Philosophy. 4 (1989): 457-81.

Copeland, B. J. Artificial Intelligence. Oxford: Blackwell, 1993.

Copi, Irving M., and Keith Burgess-Jackson. Practical Logic. New York: Macmillan, 1996.

Copp, David and David Zimmerman, ed. Morality, Reason, and Truth: New Essays in the Foundations of Ethics. Totowa, N. J.: Rowman and Allanheld, 1985.

Cummins, Robert. "Functional Analysis." Conceptual Issues in Evolutionary Biology: An Anthology, Ed. Elliot Sober. 1st. ed. Cambridge, Mass.: MIT P, 1984. 386-407.

Daly, Martin, and Margo Wilson. Homicide. Hawthorne, N. Y.: Aldine de Gruyter, 1988.

Davis, Morton D. Game Theory: A Nontechnical Introduction. New York: basic Books, 1983, 1997.

Dawkins, Richard. The Extended Phenotype: The Long Reach of the Gene. Oxford: Oxford UP, 1982.

---. "In Defense of Selfish Genes." http://www.royalinstitutephilosophy.org/articles/dawkins_genes.htm. Accessed February 2001.

Dennett, Daniel. Elbow Room. Cambridge, Mass.: MIT, 1991.

---. "Mother Nature Versus the Walking Encyclopedia: A Western Drama." In Ramsey, Stich & Rumelhardt. 21-30.

---. Darwin's Dangerous Idea: Evolution and The Meanings of Life. New York: Touchstone, 1995.

Dixit, Avinash, and Susan Skeath. Games of Strategy. New York: W. W. Norton, 1999.

Diamond, Jared. Guns, Germs, and Steel. New York: W. W. Norton, 1999.

Durkheim, Emil. The Rules of Sociological Method. Glencoe, Ill.: Free Press, 1938/1950.

Dusek, Val. "Sociobiology Sanitized: The Evolutionary Psychology and Genic Selectionism Debates." Science as Culture. 1999. <http://www.shef.ac.uk/~psych/rmy/dusek.html>

Edel, May and Abraham Edel. Anthropology and Ethics. Springfield, Ill.: Thomas, 1959.

Ellis, Bruce J. "The Evolution of Sexual Attraction: Evaluative Mechanisms in Women." Barkow et al. 267-288.

Fenton, Andrew. "Naturalized Epistemology at the end of the Twentieth Century: Moving Beyond Human Knowing." Saskatoon, Sask: WCPA, Sept. 1999.

Futuyma, Douglas J. Science on Trial: The Case for Evolution. New York: Random House, 1983.

Gauthier, David. Morals by Agreement. Oxford: Clarendon, 1986.

Goldman, Alvin I. Knowledge in a Social World. Oxford: Oxford UP, 1999.

Goodman, Nelson. Languages of Art: An Approach to a Theory of Symbols. Indianapolis: Hackett, 1976.

Haddon, Genia Pauli. "The Personal and Cultural Emergence of Yang-Feminity." To Be a Woman: The Birth of the Conscious Feminine. Ed. Connie Zweig. Los Angeles: Jeremy P. Tarcher, 1990.

Harding, Sandra. "Rethinking Standpoint Epistemology." Feminist Epistemologies. Ed. Linda Alcoff and Elizabeth Potter. New York: Routledge, 1993. 49-81.

Harris, Judith Rich. "Where is the Child's Environment? A Group Socialization Theory of Development" Psychological Review. 102 (1995): 458-89.

---. The Nurture Assumption: Why Children Turn Out the Way They Do. New York: Touchstone, 1999.

Hauptli, Bruce W. The Reasonableness of Reason: Explaining Rationality Naturalistically. Peru, Illinois: 1995.

Hornsby, Jennifer. "Collectives and Intentionality." Philosophy and Phenomenological Research. 71.2 (1997): 429-434.

Hume, David. A Treatise on Human Nature. Ernest C. Mossner, ed. London: Penguin ((1739,1740) 1969).

Humphrey, Nicholas T. "The Social Function of Intellect." Byrne and Whiten. 13-26.

Hurley, S. L. "Newcomb's Problem, Prisoners' Dilemma, and Collective Action." Synthese. 86 (1991): 173-196.

---. "A new take from Nozick on Newcomb's problem and Prisoners' Dilemma." Analysis, 54.2 (1994): 65-72.

Jolly, Alison. "Lemur Social Behavior and Primate Intelligence." Byrne and Whiten. 27-33.

---. "Primate Communication: Lies, and Ideas." Lock and Peters. 167-77.

Jones, Stephen et al. The Cambridge Encyclopedia of Human Evolution. Cambridge: CUP, 1992.

Kelley, Kevin. Out of Control: The New Biology of Machines, Social Systems, and the Economic World. Reading, Mass.: Addison Wesley, 1994.

Kymlicka, Will. Liberalism, Community, and Culture. Oxford: Oxford University Press, 1989.

---. "Liberalism." The Oxford Companion to Philosophy. Ted Honderich, ed. Oxford: OUP, 1995. 483-5.

Klein, Richard G. "Anatomy, Behavior, and Modern Human Origins." Journal of World Prehistory. 9.2 (1995). 167-195.

Kolak, Daniel, ed. The Philosophy Source. CD-ROM. Belmont, Cal.: Wadsworth, n. d.

Lewontin, Richard. Biology as Ideology: The Doctrine of DNA. Concord, Ont.: Anansi, 1991.

---. "Richard Lewontin." Accessed 28 January 2001. <http://www.biology.harvard.edu/FACULTY/Lewontin.html>

Little, Daniel. Varieties of Social Explanation: An Introduction to the Philosophy of Social Science. Boulder: Westview, 1991.

Lock, Andrew and Charles R. Peters, ed. Handbook of Human Symbolic Evolution. Oxford: Oxford UP, 1995.

Lowie, Robert H. Culture and Ethnology. New York: Peter Smith, 1917/1929.

Lykken, D. T et al. "Heritability of Interests: A Twin Study." Journal of Applied Psychology. 78.4 (1993): 649-661.

Mann, Janet. "Nurturance or Negligence: Maternal Psychology and Behavioral Preference Among Preterm Twins." Barkow, Cosmides, and Tooby, eds. 367-390.

Mellars, Paul. "Cognitive Changes and the Emergence of Modern Humans in Europe." Cambridge Archaeological Journal. 1.1 (1990): 63-70.

--- and Kathleen Gibson. Modeling the Early Human Mind. Cambridge: McDonald Institute, 1996.

Machiavelli, Niccolo. The Discourses. Trans. Leslie Walker. London: Penguin, 1970.

---. The Prince. Trans. George Bull. London: Penguin, 1981.

Megone, Christopher. "Reasoning About Rationality: Robert Nozick, The Nature of Rationality." Utilitas. 11.3 (1999): 359-374.

Mellema, Gregory. "Symbolic Value, Virtue Ethics, and the Morality of Groups." Philosophy Today. 43.3-4 (1999): 302-308.

Midgely, Mary. "Gene Juggling." Philosophy 54 (October 1979).

http://www.royalinstitutephilosophy.org/articles/midgley_gene_jugglin g.htm. Accessed February 2001.

Millikan, Ruth Garrett. "Explanation in Biopsychology." Mental Causation. Ed. John Heil and Alfred Mele. Oxford: Clarendon Press, 1993. 211-232.

Millikan, Ruth Garrett. "In Defense of Proper Functions." Philosophy of Science. 56 (1989): 288-302.

Moore, Edward C et al, ed. Writings of Charles S. Peirce. Vol. II. Indianapolis: Indiana UP, 1984.

Moser, Paul K. "Rationality, Symbolism and Evolution." International Journal of Philosophical Studies. 2.2 (1994): 287-296.

Munch, Peter A. "The Concept of 'Function' and Functional Analysis in Sociology." Philosophy of Social Science. 6 (1976): 193-213.

Noble, William and Iain Davidson. Human Evolution, Language and Mind: A Psychological and Archaeological Inquiry. Cambridge: Cambridge UP, 1996.

Nozick, Robert. Anarchy, State, and Utopia. Basic Books, 1974.

---. Philosophical Explanations. Cambridge, Mass.: Belknap, 1981.

---. The Examined Life. New York: Touchstone, 1989.

---. The Nature of Rationality. Princeton: Princeton UP, 1993.

---. Socratic Puzzles. London: Harvard UP, 1997.

---. "Invariance and Objectivity." Proceedings and Addresses of the APA. 72.2 (1997): 21-48.

Okin, Susan Muller. Justice, Gender, and the Family. Basic Books, 1989.

Pyke, Steven. Philosophers. London: Cornerhouse, 1993.

Peters, Ted. "The Problem of Symbolic Reference." Thomist. 44.1 (1980): 72-93.

Pettit, Philip. "Functional Selection and Virtual Selection." British Journal of Philosophy of Science. 47 (1996): 291-302.

Pinker, Steven. How the Mind Works. New York: W. W. Norton, 1997.

---. Words and Rules: The Ingredients of Language. New York: Basic Books, 1999.

Pinker, Steven and Paul Bloom. "Natural Language and Natural Selection." Barkow et al. 451-494.

Plantinga, Alvin. Warrant and Proper Function. New York: Oxford UP, 1993.

Primoratz, Igor. Ethics and Sex. London: Routledge, 1999.

Pranger, Robert John. Action, Symbolism, and Order.

Ramsey, William; Stephen Stich, and David Rumelhardt, eds. Philosophy and Connectionist Theory. Hillsdale, N. J.: Lawrence Erlbaum, 1991.

Ramsey, William. "Connectionism." The MIT Encyclopedia of the Cognitive Sciences. Robert A. Wilson and Frank C. Keil, eds. Cambridge, Mass.: MIT, 1999.

Rawls, John. Political Liberalism. New York: Columbia UP, 1993.

Read, Stephen J. and Lynn C. Miller, eds. Connectionist Models of Social Reasoning and Social Behavior. Mahwah, N. J.: Lawrence Erlbaum, 1998.

Rescher, Nicholas. Rationality. Oxford: Clarendon, 1988.

Rose, Steven, Leon J. Kamin, and R.,C. Lewontin. Not in Our Genes: Biology, Ideology, and Human Nature. Harmondsworth: Penguin, 1984.

Ruben, David-Hillel. "John Searle's The Construction of Social Reality." Philosophy and Phenomenological Research. 71.2 (1997): 443-447.

Ruse, Michael. E. "Evolutionary Ethics: A Phoenix Arisen." Zygon. 21 (1986): 95-112.

---. The Darwinian Paradigm: Essays on its History, Philosophy, and Religious

Implications. London: Routledge, 1993.

---. Evolutionary Naturalism: Selected Essays. London: Routledge, 1995.

---. "How Philosophy Ruined a Brilliant Scientist." Book Review. The Globe and Mail. (04/01/2000): D4.

Searle, John. Minds, Brains, and Science. London: BBC, 1984.

---. "Minds, Brains, and Programs." Introduction to Philosophy: Classical and Contemporary Readings. Ed. John Perry and Michael Bratman. Oxford: Oxford UP, 1986. 391-414.

---. "Indeterminacy, Empiricism, and the First Person." The Journal of Philosophy. 84.3 (1986): 123-146.

---. The Construction of Social Reality. New York: The Free Press, 1995.

---. "Précis of The Construction of Social Reality." Philosophy and Phenomenological Research. 71.2 (1997): 427-8.

---. "Responses to Critics of The Construction of Social Reality." Philosophy and Phenomenological Research. 71.2 (1997): 449-458.

---. Mind, Language, and Society: Philosophy in the Real World. New York: basic Books, 1998.

Sartre, Jean-Paul. Existentialism and Human Emotions. New York: Citadel, 1985.

Simpson, Jeffry A. and Douglas T. Kendrick, ed. Evolutionary Social Psychology. Mahwah, N. J.: Lawrence Erlbaum, 1997.

Skyrms, Brian. Evolution of the Social Contract. Cambridge: C UP, 1996.

Smith, Michael. The Moral Problem. Oxford: Blackwell, 1994.

Sober, Elliot. Philosophy of Biology. Boulder: Westview, 1993.

---. From a Biological Point of View: Essays in Evolutionary Philosophy. Cambridge: Cambridge UP, 1994.

Sperber, Dan. Rethinking Symbolism. Trans. Alice L. Morton. Cambridge: Cambridge UP, 1975.

Tellegen, Auke et al. "Personality Similarity in Twins Reared Apart and Together." Journal of Personality and Social Psychology. 54.6 (1988): 1031-1039.

Tooby, John and Leda Cosmides, "The Psychological Foundations of Culture." Barkow et al. 19-136.

Tuomela, Raimo. "Searle on Social Institutions." Philosophy and Phenomenological Research. 71.2 (1997): 435-441.

Underhill, P. et al. "Y Chromosome Sequence Variation and the History of Human Populations." Nature Genetics. 26 (3): 358-61.

Verkuyten, Maykel. "Symbols and Social Representations." Journal for the Theory of Social Behavior. 25.3 (1995): 263-284.

Wilson, E. O. Sociobiology. Cambridge, Mass.: Harvard UP, 1975.

---. On Human Nature. Cambridge, Mass.: Harvard UP, 1978.

Wright, Larry. "Functions." Conceptual Issues in Evolutionary Biology: An Anthology. Ed. Elliot Sober. 1st. ed. Cambridge, Mass.: MIT, 1984. 347-368.